Received: 29 October 2023 | Revised: 2 January 2024 | Accepted: 11 January 2024 | Published online: 24 January 2024

RESEARCH ARTICLE

Artificial Intelligence and Applications
2024, Vol. 00(00) 1–14
DOI: 10.47852/bonviewAIA42021969

BON VIEW PUBLISHING

Statistical Machine Learning Model for Distributed Energy Planning in Industrial Park

Chen Zhang[1,2], Xiurong Zhang[1,2,*], Xianping Wu[1,3] and Saddam Aziz[4]

[1]*College of Information and Electrical Engineering, China Agricultural University, China*

[2]*National Innovation Center for Digital Fishery, China Agricultural University, China*

[3]*State Grid Zhangzhou Power Supply Company, China*

[4]*Centre for Advances in Reliability and Safety, The Hong Kong Polytechnic University, China*

Abstract: With the advancement of industrial modernization, industrial parks have become the main body of new energy production and consumption. However, due to the large demand for energy in industrial agglomeration, the way of energy utilization is changing to sustainable. The direct connection of distributed energy resources in industrial parks, including photovoltaic (PV) power generation systems, has an important impact on its planning and operation. Furthermore, weather scenarios can have an impact on distributed PV generation, and the uncertainty in PV power output will, in turn, affect the uncertainty in industrial park planning. Therefore, this paper aims to address the issues of inaccurate prediction of distributed electricity generation during the planning period and the non-uniform distribution of energy resources such as electricity, heating, and cooling. This is achieved through the application of statistical machine learning (SML). This paper intends to incorporate the ideas of SML into the model for industrial park distributed energy random opportunity-constrained planning, aiming to resolve the problems of non-uniform distribution of distributed energy sources within the park, along with uncertainty in their outputs and high overall investment costs. The model takes the planning, construction, and operating costs of the industrial park as the objective function, uses the lost load cost to ensure the safety of the industrial park, and uses the Chebyshev's inequality probability to limit the output characteristics of distributed energy equipment. In terms of operation, the planning period is subdivided into heating period, cooling period, and transition period, and the balance of electricity, heat, and cold is considered. Finally, an actual example of an industrial park is used to verify the effectiveness of this method. Experimental validation shows that this approach can ensure safety requirements in industrial parks during the heating season, cooling season, and transitional periods by flexibly adjusting the confidence threshold. Simultaneously, it delivers significant economic benefits.

Keywords: Statistical machine learning, Bayesian generative adversarial network, weather simulation, scenario simulation, distributed energy, random chance constraint programming

1. Introduction

As the energy demand continues to grow, agricultural and industrial zones, as focal points of energy consumption, exhibit characteristics of high energy demand, centralized energy utilization, and the aggregation of multiple loads, including cooling, heating, electricity, and gas (Fu & Niu, 2023a). Traditional methods of supplying power to these cooling, heating, electricity, and gas loads separately often struggle to achieve efficient and clean energy utilization. This typically manifests as centralized power generation from conventional power plants, electricity-based cooling in the summer, boiler-based heating in the winter, or centralized heating from combined heat and power plants.

However, for industrial parks employing traditional combined cooling, heating, and power (CCHP) methods; in addition to renewable energy sources such as solar power, the choice of primary energy sources may also include the use of clean natural gas. This can provide local cooling, heating, and electricity services to end-users. Therefore, the planning of industrial parks with a focus on multi-energy integration, leveraging photovoltaic (PV) power generation and CCHP systems, supported by distributed energy storage technologies, constitutes an effective approach to address the irrational use of energy in industrial parks (Zhang et al., 2023). This not only promotes the optimal complementarity of various energy sources within the industrial park, fostering the distributed utilization of energy, but also aligns with the national calls for low-carbon development, energy efficiency, and emissions reduction.

This paper addresses the theory and methodology of stochastic planning for distributed energy resources in industrial parks, based on statistical machine learning (SML). The primary contributions of this work are as follows:

i) Addressing the challenge of large-scale and complex data types in numerical weather scenarios, where the lack of precision and diversity limits the transmission of data value to decision-makers in industrial park planning. We decided to use Bayesian generative adversarial networks (BGANs) for weather scene simulation.
ii) In response to the pressures faced by industrial parks, including uncertainty in distributed PV output, unscientific allocation of distributed energy capacities, and excessive overall investment costs, we introduce an SML-based method for random

*Corresponding author: Xiurong Zhang, College of Information and Electrical Engineering and National Innovation Center for Digital Fishery, China Agricultural University, China. Email: zhangxiurong@cau.edu.cn

opportunity-constrained planning of distributed energy in industrial parks.

This paper is structured as follows: Section 2 summarizes previous related research, and Section 3 introduces the innovative method. Case studies are presented in Section 4, and Section 5 provides the conclusion of this paper.

2. Literature Review

For a specific distributed PV power station, it often faces uncertainties arising from atmospheric factors, which encompass variables such as temperature, wind speed, cloud cover, and solar irradiance. These atmospheric elements introduce inherent unpredictability. It is worth noting that historical monitoring data already encompass factors such as angle and position. With the increase in data acquisition, various types of monitoring data exhibit the following characteristics: a combination of structured and unstructured data, discretized data from diverse collection systems, substantial data volume, and disparate data quality.

To capture stochastic processes, numerous relevant studies have been conducted. Morales et al. (2010) introduced autoregressive moving average models to renewable energy generation sites to generate spatiotemporal scenarios for a given generation profile. Yunus et al. (2016) aimed to capture the temporal dependencies and probability distributions of wind speed time series observations by improving the autoregressive integrated moving average modeling method. Hoeltgebaum et al. (2018) attempted to capture and simulate the long-term joint distribution of multivariate time series through a general autoregressive conditional score model with time-varying parameters and arbitrary non-Gaussian distributions.

Recently, machine learning (ML) theory has been increasingly applied to stochastic scenario simulation in integrated energy systems. Chen et al. (2018a) introduced generative adversarial network (GAN) methods into their research to generate scenarios using unsupervised learning, successfully capturing reliable wind power generation scenario distributions. However, this method can be challenging to train and may face issues such as gradient vanishing, and it may also produce scenarios with limited diversity. To address these challenges, Liu et al. (2019) built upon GANs by replacing the Jensen–Shannon distance with the Wasserstein distance and used Wasserstein-GAN to train renewable energy scenarios, effectively resolving the aforementioned problems. Jiang et al. (2018) further improved the performance of GAN in scenario generation by imposing Lipschitz constraints on the discriminator network.

In the works of Saatci and Wilson (2017) and Chen et al. (2018b), Bayesian formulas were integrated with genetic neural networks to ensure the simultaneous distinction and generation of renewable energy scenarios similar to historical data. This combination guarantees effective representation of the generation process of new energy, even in scenarios intentionally blending wind and solar energy data. In addition, the widespread adoption of distributed energy sources in the context of low-carbon electricity and smart industrial parks has posed challenges for the comprehensive energy system planning of such parks. Uncertainty issues can render the solutions to deterministic planning problems suboptimal or even infeasible. Due to practical considerations, uncertainty in planning problems has received significant attention (Liu, 1997). Methods for addressing uncertainty in planning problems primarily include fuzzy planning (Li et al., 2016), stochastic planning (Fu et al., 2022), and robust optimization (Li et al., 2024).

Fuzzy planning is currently limited by the subjective nature of the membership functions for fuzzy variables, often derived from experiments or personal expertise. Robust optimization, on the other hand, tends to be economically conservative as it overlooks probability distribution information, including that of distributed energy sources. In contrast, stochastic planning allows decision-makers to understand the relationship between risk and potential planning outcomes, facilitating adjustments based on real-world conditions. Consequently, it provides a better balance between the economic and security aspects of integrated energy systems in industrial parks.

Two-stage stochastic programming theory, introduced by Dantzig in 1956, divides planning problems into two stages. These stages occur before and after the realization of random variables. The first stage involves generating preliminary optimal decisions, while the second stage entails compensatory adjustments to these initial optimal decisions. In practice, two-stage stochastic programming models have proven applicable to a variety of distributed energy system issues, including design optimization (Mavromatidis et al., 2018), sizing and control of energy storage systems in integrated energy systems (Bucciarelli et al., 2018), capacity planning and energy management strategies for renewable energy components (Li et al., 2020), and enhancing the economic efficiency and system flexibility of distributed energy (Wu et al., 2020). To address uncertainty in the requirements, Huang et al. (2016) introduced a two-stage stochastic programming theory into a multi-region optimization model. They integrated decision tree methods and Monte Carlo simulation into the model, simplifying electricity demand based on node structures and determining the values and probabilities of electricity demand. This improvement helped mitigate uncertainty related to electricity demand. Fu and Zhou (2023b) considered the meteorological sensitivity of agricultural production and PV generation. They established agricultural meteorological models and energy meteorological models, proposing a new method to optimize PV greenhouse load control in collaboration with rural energy systems. This approach resulted in cost savings for greenhouse energy.

As two-stage stochastic programming theory has advanced, multi-stage stochastic programming theory has also seen gradual progress. According to multi-stage stochastic programming, uncertainty is dynamically updated, leading to mutual interactions between decision-making and uncertainty. This dynamic interaction results in more accurate and realistic outcomes. Ding et al. (2017) addressed the uncertainty associated with node injection power in power grid planning decisions. They made improvements to the multi-stage scenario tree model by incorporating uncertainty injected at nodes, which formed the basis for the multi-stage stochastic programming model. This model comprehensively accounts for future uncertainties. Hafiz et al. (2019) utilized a multi-stage stochastic programming model to optimize the problem and achieve the minimum daily purchasing cost for communities. They achieved this by formulating energy management control strategies to support decision-making to minimize costs effectively. Fu et al. (2024) proposed an optimization strategy for rural microgrids that considers adjustments to agricultural greenhouse loads. The aim is to address issues related to insufficient utilization of local renewable energy while simultaneously reducing excessive daily operational costs.

In summary, this paper addresses the challenges of large-scale, complexly distributed, low-accuracy, and non-diverse weather scenarios that are ineffective in conveying information to industrial park planning decision-makers. To overcome these issues, we

propose the use of BGAN for simulating weather scenarios. In comparison to traditional methods, this approach can acquire large-scale, high-precision, and diverse datasets of distributed PV operational scenarios. Importantly, it effectively communicates these datasets to industrial park planning decision-makers. In response to the pressures faced by industrial parks, including the uncertainty of distributed PV output, improper allocation of distributed energy capacities, and high overall investment costs, we present an SML-based method for industrial park distributed energy planning under stochastic opportunity constraints. Finally, through case studies, we validate the economic and security aspects of the balanced planning solutions. Sensitivity analyses are conducted on the performance parameters of CCHP systems and gas prices. We discuss how the uncertainties and performance parameters of distributed energy impact the overall cost of industrial park planning.

3. Proposed Method

3.1. Weather scenario simulation based on BGAN

Weather scenario data exhibit certain regularities and seasonal patterns. Traditional numerical simulation methods rely on complex probability analyses, providing strong interpretability but encountering limitations when dealing with high-dimensional data. They often yield low accuracy and can only simulate low-dimensional weather scenarios of a single category. ML algorithms possess robust autonomous learning and pattern recognition capabilities. While weather simulation outcomes can be comparable to traditional probabilistic numerical models, their internal workings are often likened to black boxes, lacking interpretability and thus lacking convincing explanations. SML combines the strengths of probabilistic models and deep learning algorithms, amalgamating the benefits of both. This amalgamation enables the simulation of high-dimensional, complex, and diverse weather scenarios while maintaining high precision. Hence, this paper utilizes BGAN to simulate and generate weather scenarios, enhancing traditional GANs through improvements in network weight parameters, loss functions, and gradient descent algorithms. From a data-physics-driven perspective, uncertainty associated with weather variables is modeled. The adversarial game in neural networks enables the neural network to learn how to generate a weather scenario dataset through a random noise distribution. Subsequently, by applying the BGAN approach to real-world scenarios, simulated weather scenarios are generated, providing concrete scenario conditions for decision-making in the context of industrial park distributed energy planning.

3.1.1. Fundamentals of GAN

According to the GAN principle (Liu et al., 2019), let the weight parameters of a generative network G be denoted as θ_G, and the weight parameters of a discriminative network D be denoted as θ_D. During the network training, the Generative network receives noise data input z_i^{noise}, which undergoes sequential sampling and training across its layers of neurons, resulting in the generation of a novel data distribution of scenarios, denoted as $G(z_i^{noise};\theta_G)$. Simultaneously, the discriminative network and generative network undergo concurrent training. The input to discriminative network comprises genuine weather data samples X, and its output discerns the category of whether x_i originates from the authentic data distribution.

3.1.2. Network weight refinement

The weight parameters of the generative network and the generated samples are initially introduced with a prior distribution as follows:

$$\theta_G \sim p_G\{\theta_G\} \tag{1}$$

$$\widetilde{x}_i = G(z_i^{noise};\theta_G) \sim p_G\{x_i\} \tag{2}$$

where θ_G represents the weight parameters of the generative network, $G(z_i^{noise};\theta_G)$ stands for the generated data scene distribution, z_i^{noise} represents the input noise, and $p_G\{\theta_G\}$ and $p_G\{x_i\}$ represent the prior distributions of weight parameters and generated samples, respectively.

Introducing the Bayesian formula into the weight parameters of the discriminative network and the generative network, iterative sampling is conducted from the conditional posterior, as follows (Saatci et al., 2017):

$$p\{\theta_D|z^{noise},X,\theta_G\} \propto \prod_{i=1}^{N_D} D(x_i;\theta_D) \times \prod_{i=1}^{m}(1-D(G(z_i^{noise};\theta_G);\theta_D)) \times p(\theta_D|\alpha_D) \tag{3}$$

$$p\{\theta_G|z^{noise},\theta_D\} \propto \left(\prod_{i=1}^{N_G} D(G(z_i^{noise};\theta_G);\theta_D)\right) \times p(\theta_G|\alpha_G) \tag{4}$$

where θ_D represents the weight parameters of the discriminative network. $p(\theta_D|\alpha_D)$ and $p(\theta_G|\alpha_G)$ represent the prior distributions of the weight parameters for the discriminative network and the generative network, respectively. α_D and α_G are hyperparameters for the weight parameters of the discriminative network and the generative network. N_D and N_G denote the numbers of input samples for the discriminative network and the generative network, respectively. $p\{\theta_D|z^{noise},X,\theta_G\}$ and $p\{\theta_G|z^{noise},\theta_D\}$ represent the posterior distributions of the weight parameters for the discriminative network and the generative network, respectively, given the known parameters X,θ_G,z^{noise}. $G(z_i^{noise};\theta_G)$ represents the weather scenarios generated by the generator under the known parameters z_i^{noise} and θ_G. $D(x_i;\theta_D)$ and $D(G(z_i^{noise};\theta_G);\theta_D)$ denote the discriminative network's outcomes for real weather scenes and generated weather scenes, respectively.

By incorporating Bayes' theorem to marginalize over the noise z_i^{noise}, a straightforward approach for handling this marginalization is through the utilization of the simple Monte Carlo method:

$$\begin{aligned} p\{\theta_G|\theta_D\} &= \int p\left\{\theta_G, z_j^{noise}|\theta_D\right\}dz \\ &= \int p\left\{\theta_G\middle|z_j^{noise},\theta_D\right\}p\left(z_j^{noise}|\theta_D\right)dz \\ &\approx \frac{1}{J_G}\sum_{j=1}^{J_G} p\left(\theta_G\middle|z_j^{noise},\theta_D\right) \end{aligned} \tag{5}$$

Similarly, we can derive

$$p\{\theta_D|\theta_G\} \approx \frac{1}{J_D}\sum_{j=1}^{J_D} p\left(\theta_D\middle|z_j^{noise},X,\theta_G\right) \tag{6}$$

where J_G and J_D represent the number of simple Monte Carlo samples taken from the generator and discriminator, respectively.

It is worth noting that when considering Equations (3) and (4) as functions of the noise z_i^{noise}, the distributions of $p\{\theta_G|z_i^{\text{noise}},\theta_D\}$ and $p(\theta_D|z_i^{\text{noise}},X,\theta_G)$ should be relatively wide because z_i^{noise} are used to generate candidate sample data. Therefore, simple Monte Carlo and each term in the intermediate steps will provide reasonable adjustments to estimate the overall marginal posterior. Through iterative sampling, it is possible, in the limit, to compute simulated weather scenario sample data from each iteration of the generative network simulation using the approximate posterior calculations for θ_G and θ_D.

3.1.3. Loss function enhancement

Wasserstein distance, also known as the Earth mover's distance, is used to measure the similarity between two distributions. It has found widespread applications in computer vision, image processing, GAN, and various other fields. The fundamental concept of the Wasserstein distance involves measuring the distance between two distributions by calculating the minimum cost required to transform one distribution into another. Its significant advantage over traditional metrics like KL divergence or JS divergence is its applicability to both discrete and continuous distributions. Furthermore, it can provide reasonable distance values even when there is ample overlap, making it capable of capturing the variability of all categories (i.e., weather scenarios) present in the training samples. This includes cases where these categories stem from the same distribution $p_{\text{data}}\{x\}$. Furthermore, the Wasserstein distance possesses a robust geometric interpretation and mathematical foundation, rendering it theoretically reliable and trustworthy. Its strong mathematical basis lends itself to the interpretability of BGAN. The mathematical formula is as follows:

$$W[D(x),G(x)]=\sup_{\theta_D}\left\{E_{x\sim p_{\text{data}}}\log D(x)+E_{x\sim p_G}[\log(1-D(x))]\right\} \tag{7}$$

where $E_{x\sim p_{\text{data}}}$ and $E_{x\sim p_G}$, respectively, represent the expectations of the distributions of real weather scenes and generated weather scenes.

3.1.4. Dynamic gradient Hamiltonian Monte Carlo descent algorithm

Hamiltonian Monte Carlo (HMC), initially proposed by Metropolis et al., is also known as the "molecular dynamics" algorithm (Brooks et al., 2011). It represents a specialized form of Markov Chain Monte Carlo algorithm. In this paper, within the posterior distribution of BGAN, HMC can assist in avoiding the complexity of the denominator in the formulas.

In essence, with knowledge of the prior distribution $p\{\theta\}$ and the data distribution $p_{\text{data}}\{X|\theta\}$, HMC can automatically facilitate sampling from the posterior distribution. Furthermore, both $p\{\theta\}$ and $p_{\text{data}}\{X|\theta\}$ are computable. The specific algorithmic procedure is outlined as follows:

- Step 1: Randomly generate a vector and set it as the initial point for iteration θ_0.
- Step 2: Perform iterative process for m iterations. At the beginning of each iteration, generate a multidimensional normal distribution vector γ_k, with the same dimension as θ, from a covariance matrix $\gamma_k\sim N(0,M)$. Here, $N(0,M)$ is referred to as the momentum.
- Step 3: During the k-th iteration, the sampled specimens generated from the previous iteration are denoted as θ_{k-1}, and the subsequent operations are executed.
- Step 4: Take the derivative of the logarithm of the objective function at the point θ_{k-1}, and update the vector γ_k to $\gamma_k+\frac{1}{2}\frac{d\log p\{\theta\}}{d\theta}|\theta_{k-1}$.
- Step 5: Utilize momentum to update the sampled specimens γ_k to $\gamma_k+\frac{1}{2}\in\frac{d\log p\{\theta\}}{d\theta}|\theta_{k-1}$.

For each update obtained in Step 3, denote the resultant θ_{k-1} as the newly generated sample θ_k, utilizing $\min\{1,r\}$ as the probabilistic selection criterion θ_k. If it falls within the specified range, θ_k becomes the new sample; otherwise, the new sample is set to θ_{k-1}. Set r as $r=\frac{p(\theta_k)p(\gamma_k)}{p(\theta_{k-1})p(\gamma_{k-1})}$.

When our objective function is defined with a posterior distribution, referencing, we can compute $\frac{d\log p\{\theta\{X\}}{d\theta}$ as $\frac{d\log p\{\theta\}}{d\theta}+\frac{d\log p\{X|\theta\}}{d\theta}-\frac{d\log p\{X\}}{d\theta}$ (Villani & Villani, 2009), which $\frac{d\log p\{X\}}{d\theta}=0$ signifies that in the calculation of the logarithmic distribution function of the posterior distribution, there is no need to evaluate intricate integrals $p\{X\}$. Similarly, in the aforementioned step g, the probability r calculation again circumvents the need for cumbersome integrals. Consequently, when employing HMC, it is only necessary to compute the prior distribution $p\{\theta\}$ and the data distribution $p_{data}\{X|\theta\}$.

3.1.5. Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP), and iterative posterior sampling

In the traditional process of using GANs for prediction, MLE is commonly employed. MLE, a method from classical frequentist statistics, estimates parameter values by maximizing the likelihood function. It is simple, intuitive, and frequently used, but its drawback lies in the absence of consideration for prior information. Therefore, in the theoretical framework of Bayesian statistics, the concept of prior knowledge is introduced, and parameter estimation is carried out by maximizing the posterior probability, known as MAP estimation. However, both MLE and MAP provide point estimates for parameters, and as a result, they may not effectively capture diverse and rich probability distributions of weather scenarios during the simulation process.

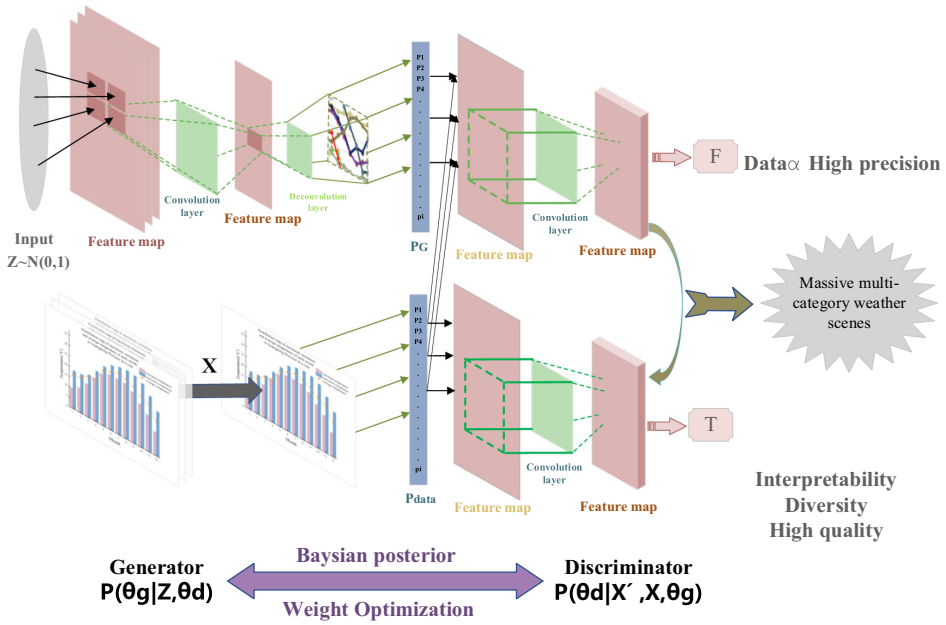
Hence, the proposed approach in this paper is based on the HMC algorithm for iterative posterior sampling. By iteratively sampling $p(\theta_D|\theta_G)$ and $p(\theta_G|\theta_D)$ at each step, approximate posterior samples for θ_G and θ_D can be obtained in the limit, thereby approximating the entire posterior distribution. The posterior distribution is extensive and multimodal, allowing for comprehensive learning of the complete data distribution characteristics of real weather scenarios. This approach enables a complete, high-precision simulation of weather scenarios.

3.1.6. Model training

For each set of weather scenario datasets $X=\{x_i\}_{i=1}^m$, there are a total of m variables, each of which follows a distribution $x_i\sim p_{\text{data}}\{x_i\}$. It should be noted that the distribution function $p_{\text{data}}\{x_i\}$ is unknown and challenging to physically model. Our objective is to draw samples from$m$noise variables $z_i^{\text{noise}}\sim p\{z_i^{\text{noise}}\}_{i=1}^m$ that follow Gaussian distributions. Using ML techniques, we aim to train a network so that the generated data samples adhere to a certain distribution p_{data}.

The proposed BGAN theory in this paper employs a convolutional neural network (CNN) as the architecture for both the discriminator and generator networks, as illustrated in Figure 1. The neural networks execute a sequence of convolutional and deconvolutional operations to extract data features. Specifically, the probability of the Bayesian posterior distribution enhances the network weights of both the generator and the discriminator. The steps for this enhancement process are as follows:

Figure 1
Schematic diagram of BGAN network architecture



- Step 1: Acquire historical samples of weather variable scenarios. The weather variables in this paper encompass outdoor ambient temperature and total solar irradiance. For each set of weather scenario datasets $\{x_i\}_{i=1}^{m}$, there are m variables, each of which follows a distribution $x_i \sim p_{\text{data}}\{x_i\}$.
- Step 2: Conduct normalization processing and generate scenario sample data X for training the learning network.
- Step 3: Initialize the generator and discriminator networks' learning rate and weight parameters, denoted as θ_{G} and θ_{D}, respectively.
- Step 4: Utilize noise z^{noise} as input for training the generator network. Initially, keep the generator network fixed. Generate a new data scenario distribution $G(z_i^{\text{noise}};\theta_{\text{G}})$. Simultaneously, train the discriminator network D along with X and update weight parameters θ_{D}. The discriminator network assesses and optimizes. Subsequently, keep the discriminator network fixed, continue training the generator network G, and update weight parameters θ_{G}. Output discernment of any class originating from the real data distribution x_i.
- Step 5: Continuously optimize both components, ultimately yielding a vast array of multi-class weather scenarios.

3.2. SML model for distributed energy planning in industrial park

According to the University of California, Berkeley, statistical learning, also known as SML, is a complex discipline. It draws upon various areas of knowledge, including probability theory, statistics, approximation theory, convex analysis, algorithmic complexity theory, and ML (Fu et al., 2020). Statistics relies on complex and rigorous mathematical reasoning, focuses on models, and emphasizes the interpretability of models. ML, on the other hand, is algorithm-oriented, prioritizes predictive results, and provides models with good controllability and scalability. The effective integration of statistical knowledge and ML makes SML a more powerful tool. Hence, in this paper, we propose an industrial park distributed energy stochastic opportunity-constrained planning model based on SML, as depicted in Figure 2.

A sound programming model is often more effective than a proficient solver. The degree of similarity between stochastic models and real-world situations significantly impacts the accuracy of solving stochastic planning problems using SML-based stochastic planning models. In this context, the paper's objective is not to enhance planning algorithms but to utilize a probabilistic constraint to ensure that the distributed energy generation characteristics satisfy a certain confidence threshold. Specifically, this is manifested by employing Chebyshev's inequality for probabilistic constraints. This approach, while reducing the conservatism of planning models, also to some extent ensures the controllability of planning and operational costs for industrial park-integrated energy systems.

3.2.1. Objective function

This paper's planning and design aim to minimize the total cost of construction and operational expenses for an industrial park's integrated distributed energy system. Additionally, it takes into account the cost incurred due to the load that cannot be supplied by each region during the heating season (Chen et al., 2021). Therefore, the objective function is as follows:

$$\min f = (C_{\text{B}} + C_{\text{O}} + C_{\text{V}}) \tag{8}$$

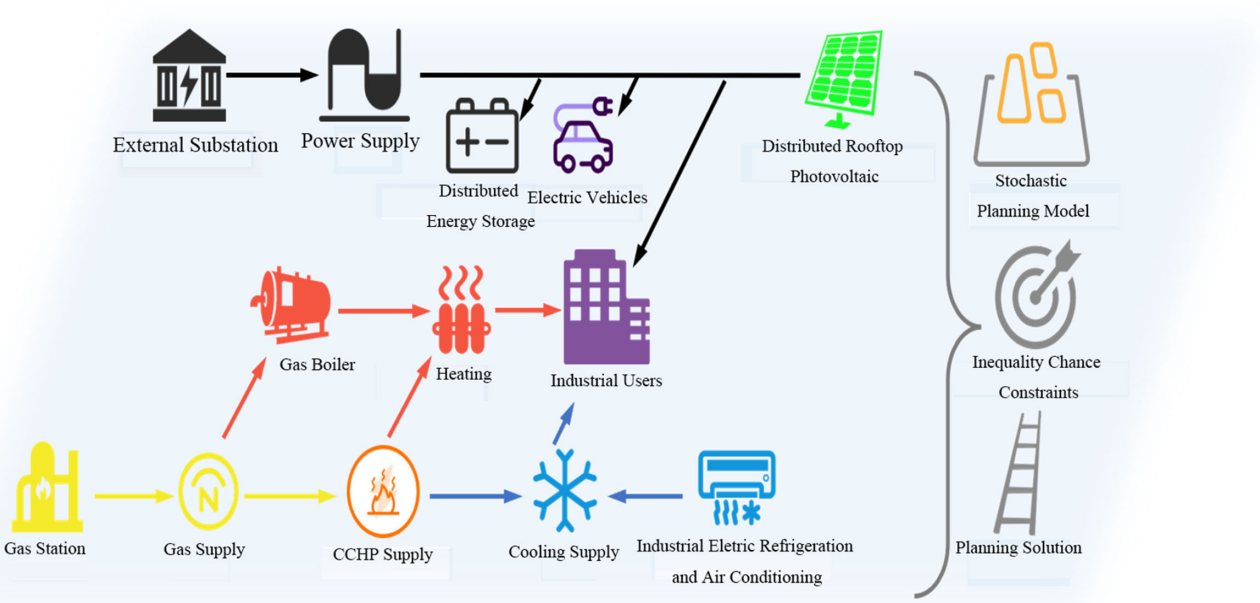
where C_{B}, C_{O}, and C_{V} represent construction cost in planning, operating cost, and unserved load cost, respectively.

a. Construction Cost in Planning C_{B}

The construction cost in planning encompasses the sum of costs for CCHP units, heating residual heat boilers, electric chillers, rooftop PV systems, and distributed energy storage systems, as shown below:

$$C_{\text{B}} = \sum_{Q\in\Psi^{\text{CCHP}}} C_Q^{\text{CCHP}} x_Q^{\text{CCHP}} + \sum_{Q\in\Psi^{\text{WH}}} C_Q^{\text{WHB}} x_Q^{\text{WHB}} + \sum_{Q\in\Psi^{\text{AC}}} C_Q^{\text{AC}} x_Q^{\text{AC}} + \sum_{Q\in\Psi^{\text{PV}}} C_Q^{\text{PV}} x_Q^{\text{PV}} + \sum_{Q\in\Psi^{\text{ESS}}} C_Q^{\text{ESS}} x_Q^{\text{ESS}} \tag{9}$$

Figure 2
Figure of the SML-based stochastic planning model for industrial parks



where C_Q^{CCHP}, C_Q^{WHB}, C_Q^{AC}, C_Q^{PV}, and C_Q^{ESS} represent the costs of the selected unit configurations Q for CCHP units, residual heat boilers, electric chillers, rooftop PV systems, and distributed energy storage systems in the industrial park areas, respectively. x_Q^{CCHP}, x_Q^{WHB}, x_Q^{PV}, and x_Q^{ESS} are 0–1 decision variables, with 1 indicating that the cost is considered. Due to their relatively small power capacity, electric chillers x_Q^{AC} are treated as continuous variables in this paper.

b. Operating Cost C_{O}

The operating costs during the planning period are divided into three categories: first, the fuel costs for CCHP; second, the electricity purchasing costs from the grid during the planning period for the industrial park; and third, the maintenance, management, and depreciation costs of distributed energy equipment. The mathematical formula is as follows:

$$C_{\mathrm{O}} = \sum_n \frac{n}{(1+i)^n}\left(\sum \mu_s\left(M^{\mathrm{GAS}} V^{\mathrm{FUEL}} + M^{\mathrm{SUB}} q_E^{\mathrm{SUB}} + M_{\mathrm{O}}\right)\right) \tag{10}$$

$$M_{\mathrm{O}} = M^{\mathrm{CCHP}} q_{\mathrm{E}}^{\mathrm{CCHP}} + M^{\mathrm{PV}} q_{\mathrm{E}}^{\mathrm{PV}} + M^{\mathrm{AC}} q_{\mathrm{E}}^{\mathrm{AC}} + M^{\mathrm{ESS}} q_{\mathrm{E}}^{\mathrm{ESS}} \tag{11}$$

$$V^{\mathrm{FUEL}} = \sum_{Q \in \Psi_j^{\mathrm{CCHP}}} V_{j,Q}^{\mathrm{CCHP}} + \sum_{Q \in \Psi_j^{\mathrm{WHB}}} V_{j,Q}^{\mathrm{WHB}} \tag{12}$$

where n represents the planning year, i is the discount rate, $\sum \frac{n}{(1+i)^n}$ denotes the total net present value of annual operating costs, M^{GAS} stands for the gas price, M^{SUB} represents the electricity price from the grid for the industrial park, M^{CCHP}, M^{PV}, M^{AC}, and M^{ESS}, respectively, denote the operation and depreciation costs of the CCHP system, distributed PV system, electric chiller system, and distributed energy storage system, with units in CNY/kW. V^{FUEL} represents the total fuel consumption within the industrial park area per unit of time, comprising two components: the fuel consumption from the residual heat boiler $V_{j,Q}^{\mathrm{WHB}}$ and the fuel consumption from the CCHP unit $V_{j,Q}^{\mathrm{CCHP}}$. $q_{\mathrm{E}}^{\mathrm{SUB}}$ represents the power demand of the original substation within the industrial park area, and μ_s signifies the proportion of each scenario within the planning period. $q_{\mathrm{E}}^{\mathrm{CCHP}}$, $q_{\mathrm{E}}^{\mathrm{PV}}$, $q_{\mathrm{E}}^{\mathrm{AC}}$, and $q_{\mathrm{E}}^{\mathrm{ESS}}$ distributions represent the power loads for the industrial park's CCHP system, distributed PV system, electric refrigeration and air conditioning, and distributed energy storage system, respectively.

c. Unserved Load Cost C_{V}

$$C_{\mathrm{V}} = M^{\mathrm{V}} \sum r \tag{13}$$

where M^{V} represents the unserved load cost coefficient, typically set at a high value to prevent load shedding during operation. r denotes the amount of load that cannot be supplied within the industrial park area.

3.2.2. Constraint conditions

1) Load Balance Constraint

Due to meteorological factors, the demand for various types of loads in the industrial park varies significantly in different seasons. This paper considers dividing the required balanced loads into three categories: electricity, cooling, and heating. During the heating season, CHP systems and gas boilers supply heat loads to user terminals, ensuring a balance between electricity and heat loads. The heating season typically spans from mid-November to mid-March. During the cooling season, cooling loads are met through electric air conditioning and lithium bromide absorption chillers. During this period, there is a balance between electricity and cooling loads. The cooling season typically spans from mid-June to mid-September. In the transitional periods, which fall outside the heating and cooling seasons, the load requirements are primarily related to electricity, as the seasonal variations in the industrial park are relatively small during these times (Chen et al., 2021).

a. Electricity Load Balance

$$r + q_{\mathrm{E}}^{\mathrm{SUB}} + q_{\mathrm{E}}^{\mathrm{PV}} = E + q_{\mathrm{E}}^{\mathrm{AC}} - q_{\mathrm{E}}^{\mathrm{CCHP}} + q_{\mathrm{E}}^{\mathrm{ESS}} \tag{14}$$

where E represents the power of electricity load within the industrial park area.

b. Thermal Load Balance

$$q_{\mathrm{H}}^{\mathrm{load}} = q_{\mathrm{H}}^{\mathrm{CCHP}} + q_{\mathrm{H}}^{\mathrm{WHB}} \tag{15}$$

$$q_{\mathrm{H}}^{\mathrm{CCHP}} = q_{\mathrm{H}}^{\mathrm{ICE}} + q_{\mathrm{H}}^{\mathrm{LMAWHC}} \tag{16}$$

where $q_{\mathrm{H}}^{\mathrm{load}}$ represents the demand for thermal load power within the industrial park area, and while $q_{\mathrm{H}}^{\mathrm{CCHP}}$ and $q_{\mathrm{H}}^{\mathrm{GL}}$, respectively, indicate the heating loads supplied by the CCHP unit and the residual heat boiler within the industrial park area. $q_{\mathrm{H}}^{\mathrm{CCHP}}$, $q_{\mathrm{H}}^{\mathrm{WHB}}$, $q_{\mathrm{H}}^{\mathrm{ICE}}$, and $q_{\mathrm{H}}^{\mathrm{LMAWHC}}$, respectively, represent the heating loads for the CCHP unit, waste heat boiler, air conditioning, and lithium bromide absorption chiller unit within the industrial park region.

c. Cooling Load Balance

$$q_{\mathrm{C}}^{\mathrm{load}} = q_{\mathrm{C}}^{\mathrm{CCHP}} + q_{\mathrm{C}}^{\mathrm{AC}} \tag{17}$$

$$q_{\mathrm{C}}^{\mathrm{CCHP}} = q_{\mathrm{C}}^{\mathrm{LMAWHC}} \tag{18}$$

where $q_{\mathrm{C}}^{\mathrm{load}}$ represents the demand for cooling load power within the industrial park area, while $q_{\mathrm{C}}^{\mathrm{CCHP}}$ and $q_{\mathrm{C}}^{\mathrm{AC}}$, respectively, indicate the cooling loads supplied by the CCHP unit and the electric chiller within the industrial park area. $q_{\mathrm{C}}^{\mathrm{CCHP}}$, $q_{\mathrm{C}}^{\mathrm{AC}}$, and $q_{\mathrm{C}}^{\mathrm{LMAWHC}}$ distributions represent the cooling loads for the CCHP unit, electric refrigeration and air conditioning, and lithium bromide absorption chiller unit within the industrial park region.

2) Distributed Energy Generation Constraint

a. Gas Internal Combustion Engine

The operational range constraint for the gas internal combustion engine in the CCHP system is

$$\sum_{Q\in\Psi^{\mathrm{CCHP}}} x_Q^{\mathrm{CCHP}} q_{min,Q}^{\mathrm{CCHP}} \le q_{\mathrm{E}}^{\mathrm{CCHP}} \le \sum_{Q\in\Psi^{\mathrm{CCHP}}} x_Q^{\mathrm{CCHP}} q_{max,Q}^{\mathrm{CCHP}} \tag{19}$$

where $q_{max,Q}^{CCHP}$ and $q_{\mathrm{min},Q}^{CCHP}$, respectively, represent the upper and lower limits of the active power output of the internal combustion engine.

b. Lithium Bromide Absorption Chiller

The construction constraint for the lithium bromide absorption chiller is

$$q_{\mathrm{H,min}}^{\mathrm{LMAWHC}} \le \mathrm{q}_{\mathrm{H}}^{\mathrm{LMAWHC}} \le q_{\mathrm{H,max}}^{\mathrm{LMAWHC}} \tag{20}$$

$$q_{\mathrm{C,min}}^{\mathrm{LMAWHC}} \le q_{\mathrm{C}}^{\mathrm{LMAWHC}} \le q_{\mathrm{C,max}}^{\mathrm{LMAWHC}} \tag{21}$$

$$q^{\mathrm{R}} \le q^{\mathrm{GAS}} + q^{\mathrm{WA}} \tag{22}$$

where $q_{\mathrm{H,max}}^{\mathrm{LMAWHC}}$ and $q_{\mathrm{C,max}}^{\mathrm{LMAWHC}}$, respectively, represent the maximum heating/cooling capacity of the lithium bromide absorption chiller, while $q_{\mathrm{H,min}}^{\mathrm{LMAWHC}}$ and $q_{\mathrm{C,min}}^{\mathrm{LMAWHC}}$, respectively, represent the minimum heating/cooling capacity. Equation (22) states that the sum of the usable heat values from the internal combustion engine's exhaust gases and cylinder liner water should be greater than or equal to the waste heat used for heating/cooling in the lithium bromide absorption chiller. q^{WA}, q^{GAS}, and q^{R}, respectively, represent the recoverable waste heat from jacket water and exhaust gases of internal combustion engines, as well as the recoverable waste heat utilized for heating/cooling by the lithium bromide absorption chiller unit.

c. Heating Residual Heat Boiler

The operational range constraint for the heating residual heat boiler is

$$\sum_{Q\in\Psi^{\mathrm{WHB}}} x_Q^{\mathrm{WHB}} q_{min,Q}^{\mathrm{WHB}} \le q_{\mathrm{H}}^{\mathrm{WHB}} \le \sum_{Q\in\Psi^{\mathrm{WHB}}} x_Q^{\mathrm{WHB}} q_{max,Q}^{\mathrm{WHB}} \tag{23}$$

where $q_{min,Q}^{\mathrm{WHB}}$ and $q_{max,Q}^{\mathrm{WHB}}$, respectively, represent the maximum and minimum heating capacity of the heating residual heat boiler.

d. Distributed Rooftop PV Generation System

For the random variable:

$$\Pr\left\{\sum_{Q\in\Psi^{\mathrm{PV}}} x_Q^{\mathrm{PV}} q_{min,Q}^{\mathrm{PV}} \le q_{\mathrm{E}}^{\mathrm{PV}} \le \sum_{Q\in\Psi^{\mathrm{PV}}} x_Q^{\mathrm{PV}} q_{max,Q}^{\mathrm{PV}}\right\} \ge \beta \tag{24}$$

where $q_{\mathrm{min},Q}^{\mathrm{PV}}$ and $q_{max,Q}^{\mathrm{PV}}$, respectively represent, the upper and lower limits of the active power outputs of the distributed rooftop PV generation system.

e. Distributed Energy Storage Model

The power limits for charging and discharging of electrical energy storage can be expressed as follows:

$$0 \le q_{\mathrm{dis,char}}^{\mathrm{ESS}}(i) \le q_{\mathrm{max}}^{\mathrm{ESS}} \tag{25}$$

$$\mathrm{SOC}_{\mathrm{min}} \le \mathrm{SOC}(i) \le \mathrm{SOC}_{\mathrm{max}} \tag{26}$$

$$q_{\mathrm{dis}}^{\mathrm{ESS}}(i)\, q_{\mathrm{char}}^{\mathrm{ESS}}(i) = 0 \tag{27}$$

where $q_{\mathrm{dis,char}}^{\mathrm{ESS}}$ represents the upper limit of the power for charging and discharging in an energy storage system, with separate limits for charging and discharging, $q_{\mathrm{dis}}^{\mathrm{ESS}}$ and $q_{\mathrm{char}}^{\mathrm{ESS}}$ represent the charging and discharging of the energy storage system, and $q_{\mathrm{max}}^{\mathrm{ESS}}$ represents the maximum power of the energy storage system. $\mathrm{SOC}_{\mathrm{min}}$ and $\mathrm{SOC}_{\mathrm{max}}$, respectively, represent the upper and lower limits of the energy storage capacity. Equation (27) represents the complementary constraint for electrical energy storage, which limits that at the same moment, energy storage can only be charged or discharged.

f. Electric Chiller System

The total investment for the electric chiller system needs to exceed the cooling load demand under extreme scenarios in the industrial park. Therefore,

$$x^{\mathrm{AC}} \ge Q^{\mathrm{AC}} \tag{28}$$

where x^{AC} and Q^{AC}, respectively, represent the total investment for the electric chiller system and the cooling load demand under extreme scenarios in the industrial park.

3.2.3. Equivalence transformation of chance constraints

It is evident that in the stochastic planning model for the industrial park's distributed energy, both the objective function and the decision variables within the constraints involve random quantities. Likely, the constraints involving random variables may not hold within a certain range of values. Therefore, this paper transforms the mathematical problem involving random variables into an opportunity-constrained planning model. This is achieved by adjusting the confidence level to ensure that the constraints are satisfied (Chen et al., 2021).

$$\begin{cases} \min \bar{f} \\ s.t.\ \Pr\{f \le \bar{f}\} \ge \alpha \\ s.t.\ \Pr\{q_{\min} \le q_E \le q_{\max}\} \ge \beta \\ \\ Equation(8)\sim(28) \end{cases} \tag{29}$$

where $\Pr\{\}$ represents the probability of an event occurring; $\bar{f}$ is the minimum value that the objective function f takes on when the confidence threshold is α; and β signifies the confidence threshold at which the constraints on CCHP system output and distributed PV output are met, with β belonging to the range [0,1]. $q_{\min} \le q_E \le q_{\max}$ represents a random variable, specifically the upper and lower constraints on the output of distributed energy resources.

In this section, we use the Chebyshev's inequality to transform the opportunity-constrained planning into a deterministic planning (Yan et al., 2021). The mathematical equations are as follows:

$$\Pr\{|q_E - E(q_E)| \le \varepsilon\} \ge 1 - \frac{E^2(q_E)}{\varepsilon^2} \tag{30}$$

where $E(q_E)$ and $E^2(q_E)$ represent the expectation and variance of the random variable, which is the distributed energy output, respectively. ε is any positive number.

Combining the upper and lower bounds constraints on distributed energy output, Formula (30) can be transformed into:

$$\Pr\{q_{\min} \le q_E \le q_{\max}\} \ge 1 - \frac{E^2(q_E) + \left[E(q_E) - \frac{q_{\max}+q_{\min}}{2}\right]^2}{\left(\frac{q_{\max}-q_{\min}}{2}\right)^2} \tag{31}$$

It should be noted that the constraints derived from the Chebyshev's inequality are, in principle, conservative and primarily involve mathematical expectations and variances (Fu et al., 2020). From Chebyshev's inequality, it can be inferred that the variance of a random variable constrains the distance of the variable itself from its expectation. As ε decreases, the accuracy of the computed result gradually increases. In this section, Chebyshev's inequality is used for constraint conditions, not the objective function. This is because if the objective function is subject to too many constraints, its economic efficiency will be reduced. For the safety constraints on the upper and lower limits of distributed energy generation, if they satisfy conservative conditions, they must also satisfy safety constraints under worst-case conditions. Conservative safety constraints, while reducing feasible solutions, do not directly compromise the overall applicability of the planning solution.

4. Experimental Results

4.1. Simulation results and analysis based on BGAN for weather scenarios

4.1.1. Training progress

The data collection for this paper mainly originates from sources such as the Guangzhou Meteorological Station (N23°10, E113°20) data, NASA data, and Meteonorm data. The scenario data cover 31 years, with a sampling point resolution of 5 min. Therefore, there are a total of 105120 × 2 × 31 points for solar radiation and temperature. Due to the nature of simulating weather scenarios using BGAN, the purpose is to have the input Gaussian white noise learn the probability distribution characteristics of real weather random variables. The output is a probability distribution that closely approximates the actual weather scenarios. Evaluation of the training results is done through statistical metrics and t-SNE, eliminating the need for cross-validation data splitting or a separate validation set to assess the model's accuracy.

This paper employs a CNN to construct the BGAN. The generator network consists of two parts. The generator for noise input employs a 5-layer deconvolutional network. The generator for input weather data samples utilizes a 5-layer network, with the first three layers being convolutional layers and the last two layers being deconvolutional layers. The discriminator network employs a 6-layer convolutional network, with the final layer being fully connected. All convolutional network filters have a size of 5 × 5. Table 1 lists the sizes used in all networks. The sigmoid activation function was applied, along with L2 regularization.

Table 1
BGAN model architecture

Generator network		
Noise input	Scene data input	Discriminator network
4 × 4 × 1024	64 × 64 × 3	64 × 64 × 3
8 × 8 × 512	32 × 32 × 128	32 × 32 × 128
16 × 16 × 256	16 × 16 × 256	16 × 16 × 256
32 × 32 × 128	32 × 32 × 128	8 × 8 × 512
64 × 64 × 3	64 × 64 × 3	4 × 4 × 1024

The convergence of the discriminator network's loss function is depicted in Figure 3. Before iteration 15200, the blue curve exhibits fluctuation and instability. However, after iteration 15200, the curve starts to converge and becomes stable. It is evident that at this point, the training of the BGAN is converging. The total number of iterations is 40255, and the final BGAN achieves satisfactory training, laying the foundation for weather scenario generation.

The judgment outputs of the discriminator network for real scene data and generated data from the generator network are illustrated in Figure 4. Initially, the blue and green curves do not overlap, exhibiting significant differences with one curve up and the other down. As the number of iterations increases, both curves fluctuate between [0.3, 0.35], start to overlap, and gradually become indistinguishable. The reason for this phenomenon is that in the initial stages of the experiment, the generator network produces samples with noticeable differences from real weather scenarios due to being fed noise input. As a result, the discriminator network can easily distinguish them. However, as the network training progresses, the generated samples gradually become more realistic, narrowing the gap between them and real weather scenarios. When the number of

Figure 3
Discriminator network loss function plot

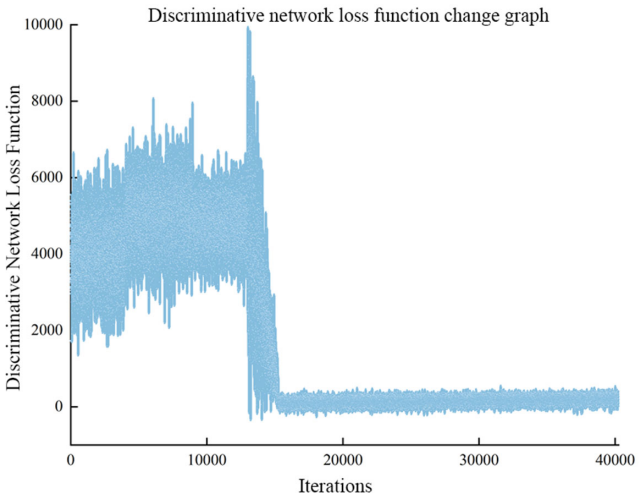
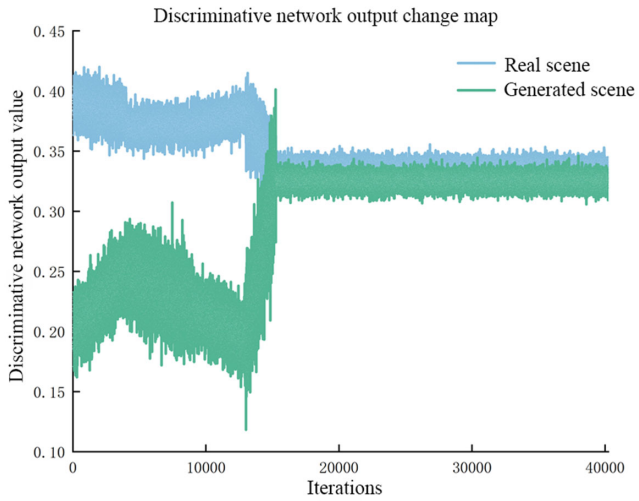


Figure 4
Discriminator network output plot



iterations reaches around 16000, the real and generated scenarios become indistinguishable from the discriminator network. The loss function of the discriminative network, which is the Wasserstein distance, gradually converges to near zero. Once the network training reaches a stable state, the simulated weather scenarios become diverse and abundant, resulting in a large collection of varied weather scenario data.

4.1.2. Validation of effectiveness

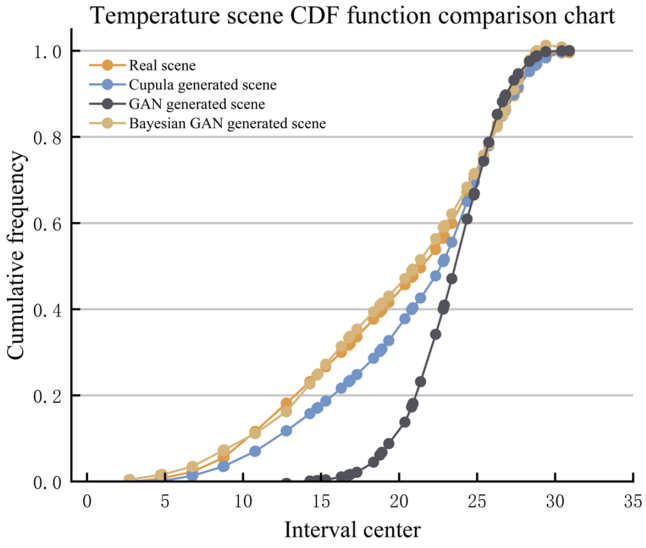
To demonstrate the effectiveness of BGAN in simulating weather scenarios, this paper utilized real solar radiation and temperature data. Weather scenes for 1 year were randomly selected from both training and testing datasets. These scenarios were then compared with statistical metrics based on copula (Wang et al., 2020), GAN (Zhu et al., 2022), and Time-series Dense Encoder (Das et al., 2023) theories. The comparison results are shown in Table 2. BGAN generated the lowest values in MAE and RMSE measurements, indicating high simulation accuracy and minimal simulation error. Furthermore, the highest R2 result for BGAN suggests strong interpretability and applicability to weather scenario datasets. These results validate the accuracy and effectiveness of the proposed BGAN algorithm in simulating weather scenarios.

Table 2
Comparison of statistical metrics for different simulation methods

Simulation methods	MAE	RMSE	R2
Copula	12.3127	12.2482	1.3577
GAN	10.1843	11.1931	1.3082
TiDE	9.8215	11.3339	1.2118
BGAN	9.7658	10.6640	1.7444

Using historical scene data based on actual inputs, this paper also compares the cumulative distribution functions of copula-based, GAN-generated, and BGAN-generated scenarios from a probability distribution perspective. This comparison is depicted in Figures 5 and 6. The blue and red curves show similar trends across the entire interval, while the green and purple curves exhibit consistency only in the final segment of the interval. From this analysis, it can be concluded that copula-based and GAN-generated scenarios capture the local features of temperature and solar radiation scenarios, while BGAN maps the overall characteristics of real weather scenarios. As a result, the performance of BGAN is superior to copula and GAN methods.

Figure 5
Comparison of weather temperature based on real, GAN, and BGAN



4.1.3. Diversity validation

In a single plot, the visualization of all real weather data sequences, along with copula, GAN, and BGAN-generated scenario data, using t-SNE distribution is shown in Figures 7 and 8. In the figures, the purple dots are more concentrated compared to the blue, green, and yellow dots. There are more clusters formed by the purple dots, and they significantly overlap with the red dot region. The coverage is the highest, and the distance is very close.

Figure 6
Comparison of solar radiation based on real, GAN, and BGAN

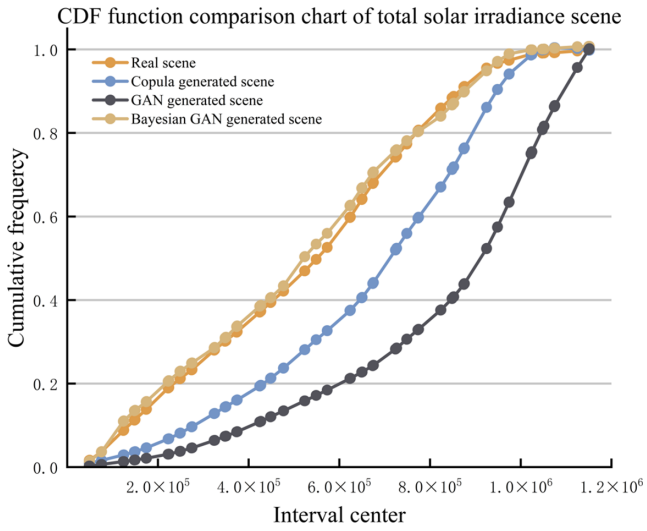


Figure 7
Dimensionality reduction visualization of weather temperature

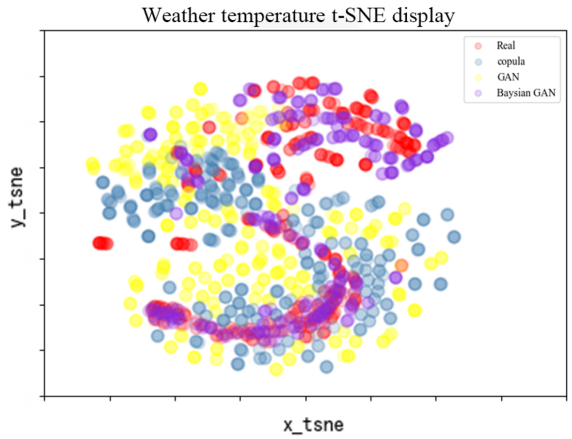
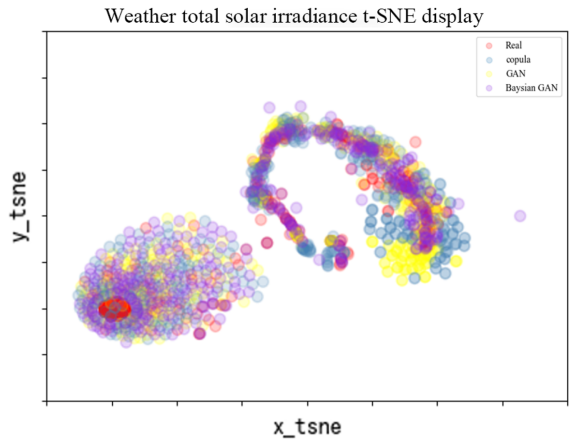


Figure 8
Dimensionality reduction visualization of weather scenarios



This indicates that the diversity of scenario results from BGAN is enhanced, and their similarity to real scenarios is higher than the scenario results from copula and GAN.

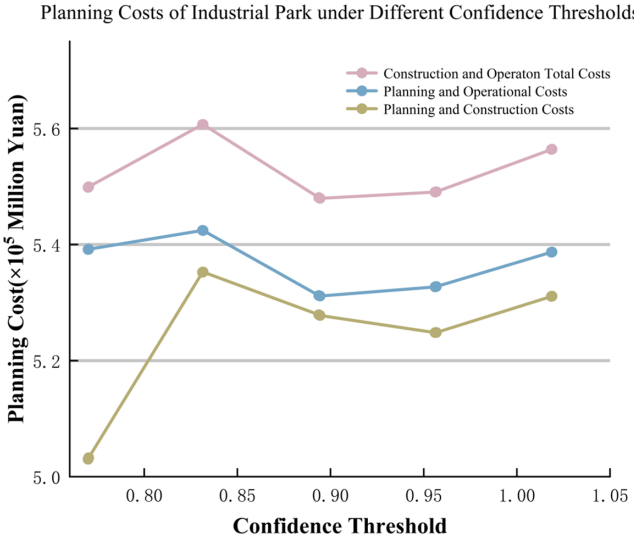
4.2. Results of chance-constrained planning

4.2.1. Planning results

This study takes the Mingzhu Industrial Park in Conghua, Guangzhou as an example. The lower heating value of fuel is 32.967 MJ/m^3 with a price of 3.23 CNY/m^3. The heating coefficient of the absorption chiller is 0.9. The average price of electricity from the external grid in the industrial park is 0.9923 CNY/kWh. The total planning period is set to 10 years, and the annual operating cost discount rate is set to 5%. The cost coefficient of load shedding is 1×10^6 CNY/MW.

Within the planning model, all constraints on the output characteristics of distributed energy resources are set at the same confidence level. The stochastic planning costs at different confidence levels are calculated, as depicted in Figure 9, with specific configuration cost results presented in Table 3.

Figure 9
Cost of integrated energy system planning for industrial park at different confidence levels



From Figure 9, it can be observed that as the confidence threshold increases, construction and operation of the industrial park-integrated energy system exhibit a trend of initially increasing, then decreasing, and then increasing again as the confidence threshold increases. When the confidence threshold is 0.90, the lowest total cost for the construction and operation planning of the industrial park's comprehensive energy system is 5.479×10^6 million CNY. From Table 3, it can be observed that in terms of planning and construction, when the confidence level is 0.95, the lowest construction cost is 1.63×10^5 million CNY. In terms of operational costs, the expenditure on external electricity procurement for the park has decreased by 5.91%. This indicates that the utilization of all distributed energy equipment increases with the rise of the confidence threshold. The CCHP system's cooling function can replace a portion of electric air conditioning, and its heating function can substitute for a portion of gas boiler heating. The distributed rooftop PV system and

Table 3
Comparison of stochastic planning results table (Unit: ten thousand CNY)

Confidence level	Planning and construction cost						Strategic planning of operational costs					Planning, construction, and total cost of operation	Cost of unloaded operation
	CCHP	Heating waste heat boilers	Distributed energy storage equipment	Distributed photovoltaic system	Electric refrigeration and air conditioning system	Total	CCHP	External procurement of electricity	District heating residual heat boiler	Operation and maintenance of distributed energy equipment	Total		
0.8	60000	60003	840	4264	42185	167292	263355	2900063	324628	1903691	5391737	5559039	1.36×10^9
0.85	71820	68576	840	4264	42185	187685	315218	2787269	378893	1937529	5418909	5606594	1.44×10^9
0.9	71820	48573	420	6396	42185	169394	315218	2728804	310158	1955069	5309249	5478643	1.29×10^9
0.95	71820	40000	840	8528	42185	163373	315218	2765996	302679	1943911	5327804	5491177	1.33×10^9
1	71820	54287	840	8528	42185	177660	315218	2845735	308934	1919989	5389876	5567536	1.31×10^9

distributed energy storage also contribute to electricity generation and storage to some extent. In terms of operational planning, when the confidence level is 0.90, the lowest operational cost is 5.31×10^6 million CNY. As the confidence level increases from 0.90 to 1.0, the operating costs have increased by 1.50%. This is due to the increased electrical, heating, and cooling load levels that the industrial park needs to meet. Therefore, as the total capacity of the integrated energy system continues to increase, it imposes excessive pressure on the business expenses during the planning period.

Regarding the results of the stochastic chance-constrained planning outlined above, the following conclusions can be drawn:

i). The industrial park's investment, planning, construction, and operational costs exhibit an initial increase followed by a decrease as the confidence threshold increases. This phenomenon can be attributed to the alignment between the confidence threshold and the load balance requirements of the system. Higher confidence thresholds correspond to more stringent load balance requirements, leading to increased costs. The total cost reaches its minimum value when the confidence threshold is 0.9.

ii). In the planning and construction of industrial parks, the cost of idle operation reflects the reliability of the power supply system. By adjusting the confidence threshold, the system ensures that each complex scenario involving heating periods, cooling periods, and transition periods meets both economic and safety requirements during operation.

iii). From a national perspective, the improvement of integrated energy configuration efficiency aims to avoid energy wastage. However, for energy supply providers in park construction, considering energy configuration efficiency ultimately serves the purpose of return on investment. According to Fu et al. (2020) and Shen et al. (2016), the overall cost of park planning and operation has decreased. This reduction reflects the benefits brought about by using probability constraints to restrict the output characteristics of distributed energy to meet confidence thresholds, impacting equipment selection and capacity configuration. It signifies an improvement in energy configuration efficiency.

4.2.2. Sensitivity analysis

For the comprehensive energy system in the industrial park, the performance fluctuations of the CCHP system's parameters and variations in gas prices will impact the economic feasibility of the park's planning results. Therefore, in this section, under the current stochastic planning outcomes, sensitivity analysis is conducted on various parameters. Specifically, the impact of a 10% fluctuation in both directions for each parameter is discussed at a confidence level of 0.9. This analysis aims to assess the influence on the industrial park's planning and construction costs, operational planning costs, and overall costs.

1) Results of Sensitivity Analysis on CCHP Unit Performance Parameters

Sensitivity analysis was conducted on the performance parameters of the CCHP unit, namely α^{ICE}, α^{GAS}, α^{WA}, β^{ICE}, β^{GAS}, and β^{WA}. The results are presented in Table 4.

This table shows that the parameter with the most significant impact on the planning results is the linear coefficient of CCHP electricity generation output, denoted as α^{GAS} and α^{GAS}. When these parameters are decreased by 10%, the 10-year planning, construction, and operational total cost for the industrial park amount to

540.26 billion CNY, resulting in an investment cost decrease of approximately 1.39%.

Table 4
Sensitivity analysis of CCHP unit performance parameters (Unit: 100 million CNY)

Parameter variations	Construction investment	Operating costs	Total costs	Fluctuation percentage (%)
0	16.94	530.92	547.86	0
$\alpha^{ICE}+10\%$	16.98	535.36	552.34	0.82
$\alpha^{ICE}-10\%$	16.86	527.22	544.08	−0.69
$\beta^{ICE}+10\%$	16.92	531.24	548.16	0.05
$\beta^{ICE}-10\%$	16.93	523.56	540.49	−1.35
$\alpha^{GAS}+10\%$	16.89	533.42	550.38	0.46
$\alpha^{GAS}-10\%$	16.96	523.37	540.26	−1.39
$\beta^{GAS}+10\%$	16.93	523.88	540.81	−1.29
$\beta^{GAS}-10\%$	16.93	531.91	548.84	0.18
$\alpha^{WA}+10\%$	16.9	530.52	547.42	−0.08
$\alpha^{WA}-10\%$	16.95	524.27	541.22	−1.21
$\beta^{WA}+10\%$	16.92	530.87	547.79	−0.01
$\beta^{WA}-10\%$	16.93	523.92	540.85	−1.28

2) Results of Sensitivity Analysis on Gas Price Parameter

In this study, the initial purchase price of gas was set at 3.23 CNY/m^3. Due to the significant fluctuations in gas prices, a sensitivity analysis was conducted on gas prices. The results are presented in Table 5. As observed from the table, gas prices also considerably impact the industrial park's overall economic feasibility. A 20% decrease in gas price can lead to a 0.36% reduction in overall costs.

Table 5
Sensitivity analysis of gas price parameter (Unit: 100 million CNY)

Parameter variation	Planning and construction investment	Operating costs	Total costs	Fluctuation percentage (%)
None	16.94	530.92	547.86	0
$M^{GAS}+10\%$	16.95	533.54	550.47	0.48
$M^{GAS}+20\%$	17.04	536.41	553.45	1.02
$M^{GAS}-10\%$	16.93	530.26	547.19	−0.12
$M^{GAS}-20\%$	16.91	528.97	545.88	−0.36

5. Conclusion

Energy and the environment have become bottlenecks constraining the sustainable development of the national economy. Industrial parks are concentrated areas of energy consumption, leading to a transformation in energy utilization toward refinement, decentralization, and sustainability, in response to the high demand for energy such as cooling, heating, and electricity from clustered industries. However, the output of distributed energy is influenced by weather conditions, exhibiting strong randomness and volatility. Moreover, with abundant and diverse monitoring data, potentially valuable operational scenario information may not be effectively conveyed to the park planning decision-makers. This can result in an imprudent allocation of distributed energy capacities within the integrated energy system, thereby subjecting the industrial park's power balance to significant pressure and posing new challenges to energy reliability and economic viability. Therefore, this study comprehensively considers the energy demands for electricity, heating, and cooling loads during the industrial park planning process, along with the forecast of renewable energy. It applies the principles of SML to the process of random planning for distributed energy. As a result, a planning method with stochastic opportunities is proposed. The detailed conclusions are as follows:

(1) A method is proposed for simulating and generating weather scenarios using BGAN. In contrast to traditional methods for simulating weather scenes, this paper introduces the principles of SML. Leveraging the characteristics of the posterior distribution in Bayesian theory, the method optimizes the weight parameters of the GAN using the posterior distribution, obtaining a probability distribution that better approximates real weather scenarios. The BGAN method takes advantage of the feedforward properties of ML neural networks. Its operability depends on the characteristics of the data model, making it easy to operate without the need for sampling or manual annotation of data, ensuring computational efficiency in solving SML problems. The improvement of the genetic algorithm based on complete probability inference is a fusion of probability theory and ML. During training, Wasserstein distance is used to assess the distance between the probability distribution $p_G\{x\}$ and real data $p_{data}\{x\}$, ensuring the representativeness of all-weather scenario categories during the training process. The dynamic gradient HMC algorithm is employed for network training, replacing the traditional stochastic gradient descent algorithm. This simplifies the training process while ensuring the stability of training convergence. Simulation results indicate that the BGAN method exhibits higher accuracy and partitioning performance. The method effectively conveys accurate and comprehensive scenario results to planning decision-makers, meeting the engineering requirements of distributed energy planning considering SML in industrial parks.
(2) A distributed energy stochastic planning model for industrial parks considering SML has been proposed. The model analyzes the construction and operational costs of industrial park planning. The considered energy supply alternatives include CCHP systems, district heating residual heat boilers, distributed rooftop PV systems, electric refrigeration and air conditioning systems, and distributed energy storage. In the model, the complex operational scenarios of the integrated energy system are divided into heating, cooling, and transition periods during the planning phase, considering the balance of electricity, heating, and cooling. The Chebyshev's inequality is employed to probabilistically constrain the output characteristics of distributed energy devices to meet a certain confidence threshold.

Simulation results demonstrate that with an increasing confidence threshold, the total cost of planning, construction, and operation exhibits an initial increase followed by a decrease and then an increase again. When the confidence threshold is set at 0.90, the construction and operational planning total cost for the industrial park's comprehensive energy system is minimized, yielding the optimal economic benefit. All three scenarios' economic and safety requirements—heating, cooling, and transition—are met by flexibly adjusting the confidence threshold during system operation. Additionally, CCHP unit performance parameters and gas prices also impact the park's planning outcomes, necessitating careful consideration of these parameters during unit selection.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in [repository name e.g., "figshare"] at http://doi.org/10.5281/zenodo.10609094

References

Brooks, S., Gelman, A., Jones, G., & Meng, X. L. (2011). *Handbook of Markov chain Monte Carlo*. USA: CRC Press.

Bucciarelli, M., Paoletti, S., & Vicino, A. (2018). Optimal sizing of energy storage systems under uncertain demand and generation. *Applied Energy*, *225*, 611–621. https://doi.org/10.1016/j.apenergy.2018.03.153

Chen, Y., Li, P., & Zhang, B. (2018b). Bayesian renewables scenario generation via deep generative networks. *52nd Annual Conference on Information Sciences and Systems*, 1–6.

Chen, Y., Wang, X., & Zhang, B. (2018a). An unsupervised deep learning approach for scenario forecasts. *Power Systems Computation Conference,* 1–7. https://doi.org/10.23919/PSCC.2018.8442500

Chen, Z., Gao, Z., Chen, J., Wu, X., Fu, X., & Chen, X. (2021). Research on cooperative planning of an integrated energy system considering uncertainty. *Power System Protection and Control*, *49*(8), 32–40.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., & Yu, R. (2023). Long-term forecasting with TiDE: Time-series dense encoder. *arXiv Preprint:2304.08424*. https://doi.org/10.48550/arXiv.2304.08424

Ding, T., Li, C., Hu, Y., & Bie, Z. (2017). Multi-stage stochastic programming for power system planning considering nonanticipative constraints. *Power System Technology*, *41*, 3566–3573.

Fu, X., Guo, Q., & Sun, H. (2020). Statistical machine learning model for stochastic optimal planning of distribution networks considering a dynamic correlation and dimension reduction. *IEEE Transactions on Smart Grid*, *11*(4), 2904–2917. https://doi.org/10.1109/TSG.2020.2974021

Fu, X., & Niu, H. (2023a). Key technologies and applications of agricultural energy internet for agricultural planting and fisheries industry. *Information Processing in Agriculture*, *10*(3), 416–437. https://doi.org/10.1016/j.inpa.2022.10.004

Fu, X., Wei, Z., Sun, H., & Zhang, Y. (2024). Agri-energy-environment synergy-based distributed energy planning in rural areas. *IEEE Transactions on Smart Grid*.

Fu, X., Wu, X., Zhang, C., Fan, S., & Liu, N. (2022). Planning of distributed renewable energy systems under uncertainty based on statistical machine learning. *Protection and Control of Modern Power Systems*, *7*(4), 1–27.

Fu, X., & Zhou, Y. (2023b). Collaborative optimization of PV greenhouses and clean energy systems in rural areas. *IEEE Transactions on Sustainable Energy*, *14*(1), 642–656. https://doi.org/10.1109/TSTE.2022.3223684

Hafiz, F., de Queiroz, A. R., Fajri, P., & Husain, I. (2019). Energy management and optimal storage sizing for a shared community: A multi-stage stochastic programming approach. *Applied Energy*, *236*, 42–54. https://doi.org/10.1016/j.apenergy.2018.11.080

Hoeltgebaum, H., Fernandes, C., & Street, A. (2018). Generating joint scenarios for renewable generation: The case for non-Gaussian models with time-varying parameters. *IEEE Transactions on Power Systems*, *33*(6), 7011–7019. https://doi.org/10.1109/TPWRS.2018.2838050

Huang, Y. H., Wu, J. H., & Hsu, Y. J. (2016). Two-stage stochastic programming model for the regional-scale electricity planning under demand uncertainty. *Energy*, *116*, 1145–1157. https://doi.org/10.1016/j.energy.2016.09.112

Jiang, C., Mao, Y., Chai, Y., Yu, M., & Tao, S. (2018). Scenario generation for wind power using improved generative adversarial networks. *IEEE Access*, *6*, 62193–62203. https://doi.org/10.1109/ACCESS.2018.2875936

Li, L., Liu, M., Cheng, G., & Liu, Q. (2016). Win-Win service policy based on fuzzy stochastic programming in E-MRO. *Computer Integrated Manufacturing Systems*, *22*(1), 70–87.

Li, R., Guo, S., Yang, Y., & Liu, D. (2020). Optimal sizing of wind/concentrated solar plant/electric heater hybrid renewable energy system based on two-stage stochastic programming. *Energy*, *209*, 118472. https://doi.org/10.1016/j.energy.2020.118472

Li, X., Li, G., Li, X., Li, X., & Li, H. (2024). Robust planning method for energy storage station considering wind power uncertainty and battery loss. *Proceedings of the CSU-EPSA*. https://doi.org/10.19635/j.cnki.csu-epsa.001209

Liu, B. (1997). Dependent-chance programming: A class of stochastic optimization. *Computers & Mathematics with Applications*, *34*(12), 89–104.

Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., & Wang, Z. (2019). Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering*, *5*(1), 156–163. https://doi.org/10.1016/j.eng.2018.11.018

Mavromatidis, G., Orehounig, K., & Carmeliet, J. (2018). Design of distributed energy systems under uncertainty: A two-stage stochastic programming approach. *Applied Energy*, *222*, 932–950. https://doi.org/10.1016/j.apenergy.2018.04.019

Morales, J. M., Minguez, R., & Conejo, A. J. (2010). A methodology to generate statistically dependent wind speed scenarios. *Applied Energy*, *87*(3), 843–855. https://doi.org/10.1016/j.apenergy.2009.09.022

Saatci, Y., & Wilson, A. G. (2017). Bayesian GAN. *31st Conference on Neural Information Processing Systems,* 1–10.

Shen, X., Shahidehpour, M., Zhu, S., Han, Y., & Zheng, J. (2016). Multi-stage planning of active distribution networks considering the co-optimization of operation strategies. *IEEE Transactions on Smart Grid*, *9*(2), 1425–1433. https://doi.org/10.1109/TSG.2016.2591586

Villani, C., & Villani, C. (2009). Stability of optimal transport. *Optimal Transport: Old and New*, 773–793.

Wang, Z., Wang, W., Liu, C., & Wang, B. (2020). Forecasted scenarios of regional wind farms based on regular vine copulas. *Journal of Modern Power Systems and Clean Energy*, *8*(1), 77–85. https://doi.org/10.35833/MPCE.2017.000570

Wu, D., Ma, X., Huang, S., Fu, T., & Balducci, P. (2020). Stochastic optimal sizing of distributed energy resources for a cost-effective and resilient Microgrid. *Energy*, *198*, 117284. https://doi.org/10.1016/j.energy.2020.117284

Yan, J., Ye, R., Zhong, H., & Jiang, X. (2021). Twice labels number estimation algorithm based on Gaussian fitting and Chebyshev inequality. *Journal of Electronics & Information Technology*, *43*(7), 1893–1899. http://dx.doi.org/10.11999/JEIT200209

Yunus, K., Thiringer, T., & Chen, P. (2016). ARIMA-based frequency-decomposed modeling of wind speed time series. *IEEE Transactions on Power Systems*, *31*(4), 2546–2556. https://doi.org/10.1109/TPWRS.2015.2468586

Zhang, X., Fu, X., Xue, Y., Chang, X., & Bai, X. (2023). A review on basic theory and technology of agricultural energy internet. *IET Renewable Power Generation*. https://doi.org/10.1049/rpg2.12808

Zhu, L., Li, W., Wang, Q., & Zhang, X. (2022). Wind farms-green hydrogen energy storage system capacity sizing method based on corrected-conditional generative adversarial network. *Transactions of China Electrotechnical Society, 39*(3), 714–730. https://doi.org/10.19595/j.cnki.1000-6753.tces.222009

How to Cite: Zhang, C., Zhang, X., Wu, X., & Aziz, S. (2024). Statistical Machine Learning Model for Distributed Energy Planning in Industrial Park. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA42021969