Protection and Control of Modern Power Systems

ORIGINAL RESEARCH

Open Access

Statistical machine learning model for capacitor planning considering uncertainties in photovoltaic power

Xueqian Fu*

Abstract

New energy integration and flexible demand response make smart grid operation scenarios complex and changeable, which bring challenges to network planning. If every possible scenario is considered, the solution to the planning can become extremely time-consuming and difficult. This paper introduces statistical machine learning (SML) techniques to carry out multi-scenario based probabilistic power flow calculations and describes their application to the stochastic planning of distribution networks. The proposed SML includes linear regression, probability distribution, Markov chain, isoprobabilistic transformation, maximum likelihood estimator, stochastic response surface and center point method. Based on the above SML model, capricious weather, photovoltaic power generation, thermal load, power flow and uncertainty programming are simulated. Taking a 33-bus distribution system as an example, this paper compares the stochastic planning model based on SML with the traditional models published in the literature. The results verify that the proposed model greatly improves planning performance while meeting accuracy requirements. The case study also considers a realistic power distribution system operating under stressed conditions.

Keywords: Uncertainty, Statistical machine learning, Stochastic programming, Renewable energy

1 Introduction

The optimal placement of distributed energy resources (DERs) and capacitor banks is an important issue in power systems. Nondeterministic characteristics of loads and DERs are important challenges for the economic and safe operation of power grids, and will greatly affect distribution network planning * MERGEFORMAT [1]. To characterize the nondeterministic characteristics of power flows, the interval power flow is an effective method. In practical systems, uncertainty brings challenges to power grid optimization. Mathematically, the interval model of power grid uncertainty faces the nonconvex nonlinear programming problem, known to be NP-hard. Energy storage allocation has become a popular method to solve uncertainty optimization problems of power grids * MERGEFORMAT [2]. An optimizing-scenario model is presented to handle the uncertain power flow problem in [3], while power flow calculations within a nonlinear programming algorithm require advanced metering infrastructure to collect smart meter data [4]. A static equivalent method is proposed to meet the optimization requirements of optimal reactive power flow using measurements in [5], whereas in [6], a static equivalent model for gas networks is proposed such that electricity–gas co-optimization becomes feasible in mathematics. Stochastic planning of distribution networks not only deals with the stochastic optimization operation described in the above literature but also pursues annual performance from the perspective of economy and technology.

To achieve optimal planning for distribution networks, uncertainty programming models are necessary, considering the uncertainties in loads and DERs. Reference [7] presents an uncertainty programming model for optimal

*Correspondence: fuxueqian@cau.edu.cn
College of Information and Electrical Engineering, China Agricultural University, 17 Qinghua Donglu, Beijing 100083, People's Republic of China

planning of plug-in electric vehicle charging stations, whereas planned energy storage based on photovoltaic (PV) correction is presented in * MERGEFORMAT [8], which analysed the economic value of energy storage. To improve frequency stability, it is suggested that wind power frequency regulation should be predicted * MERGEFORMAT [9]. An optimal planning strategy is formulated to make full use of the fast-response capability of DERs in [10], while to provide reliable planning results for microgrids, not only the stochastic nature of DERs but also the operational criteria of each power apparatus should be considered [11]. In conclusion, the stochastic programming model based on the probability distribution function has become the main method for uncertainty planning of distribution networks.

It is common that probabilistic power flow (PPF) results are available for power system planning [12]. However, PPF theory faces some difficult problems. Specifically, the requirements of PPF algorithms include being able to deal with the nonlinear correlations between new energies and random loads, and not only numerical characteristics but also the probability density function (PDF) and cumulative probability distribution function (CPDF). PPF algorithms should ensure the estimation accuracy and improve the efficiency of calculation.

Based on our previous work, this paper studies the application of probability, statistics and PPF theories to the problem of distribution network planning, subjecting it to uncertainties in random loads and PV generation. First, the combination of chance-constrained functions and particle swarm optimization (PSO) is used to solve the chance-constrained stochastic programming model considering PV uncertainty [13]. Second, the minimum load rate is considered to improve the classic loss factor method for estimating energy loss, which is an important index for the planning of distributed generation in distribution networks [14]. Third, PPF calculation methods are presented, considering the correlation and uncertainty of new energy sources in power systems [15] and integrated energy systems [16]. Finally, PPF is used to build a stochastic power system planning model as in [17].

The academic viewpoints of this paper are as follows. For power system problems with clear physical concepts and models, the data-driven method is unnecessary, while the black box may not lead to a better effect. Machine learning technology can be introduced to solve the problem of distribution network reinforcement planning considering nonlinear stochastic programming. This has a detrimental effect on the model-driven method. The explicable character of machine learning is of paramount importance in the field of artificial intelligence (AI) techniques in power systems. 'Explicit' and 'faithful' are two keys to the explainability of AI. Explicit stands for how many intersections exist between an explanation and the comprehension ability of a given group of people. The clearer the explanation is, the greater the intersections are. Faithfulness reflects the correctness of the explanation, i.e., to what extent the explanation reveals the real mechanism of the AI system. Statistical machine learning (SML) makes full use of the explanation of mathematical statistics, and this can improve the explanation of machine learning and break through the obstacles of AI application in distribution network planning. This is the motivation for the current paper.

The potential benefits deriving from the application of the proposed method can be outlined as follows.

(1) Deterministic planning cannot solve the uncertainty problems of new energy distribution networks. Robust optimization is good at dispatching and can ensure the security of the power grid. From the perspective of mathematical programming, it can obtain the maximum economy in probability under the premise of high probability security using probabilistic planning.
(2) Power distribution system planning is not characterized by strict time constraints, but it can significantly improve the feasibility of complex optimization by greatly reducing the calculation time while ensuring the accuracy of the planning model.
(3) The planning of distribution networks does not simply depend on the results of power flow or PPF.

It has been proven that probability theory and machine learning are effective methods for simulating new energy scenarios. However, the existing methods do not consider the seasonal differences of random new-energy output. Probabilistic power flow is considered to be an effective method for uncertainty analysis, but its use for uncertainty planning has not been studied. This paper presents a methodology based on statistical machine learning in power distribution networks. It focuses on the context of active distribution networks subject to uncertainties due to the large penetration of distributed renewable generation.

The main contributions of the paper can be summarized as follows.

(1) A SML-based capricious weather model is proposed, which considers not only uncertainty but also seasonality. Such a model is novel and has significance for modelling renewable energies. Based on the proposed weather model, the uncertainty simulation of annual PV power generation and cooling load is realized.

(2) A fast calculation method is proposed to analyse the uncertainty of renewable energy systems, instead of a power flow calculation based on the maximum likelihood approach, singular value decomposition, and the stochastic response surface method (SRSM).
(3) A novel probabilistic programming model for capacitor planning, one which considers uncertainties in PV generation, is proposed, and the probability information of probabilistic power flow is converted into constraint information of planning models using the central point method.

2 Problem description

Different from the traditional passive distribution networks, modern active distribution networks may contain a high proportion of distributed PV generation. Because of the randomness of user behavior, heating, ventilation and air conditioning (HVAC) loads are uncertain. Consequently, random PV output and electricity consumption behaviors bring bilateral uncertainty to the analysis and planning of distribution networks, as shown in Fig. 1.

If the uncertainty is not considered and the deterministic model is used to plan the distribution networks, the planning effect may not be optimal or even acceptable. Stochastic programming theory can be used to solve the uncertainty planning problem, while deterministic power flow (DPF) is not adequate for stochastic programming. In the solution of power system stochastic optimization problems, if there are insufficient scenarios, the uncertainty will not be described accurately. As a result, the quality of the optimal solution will be harmed and the risk to power system operation may also increase. In contrast, if the number of classic scenarios is too large, although the accuracy of the solution can be guaranteed, the computational complexity of stochastic optimization will increase dramatically. In addition, as the efficiency of the solution is reduced, the problem may even become more difficult to solve.

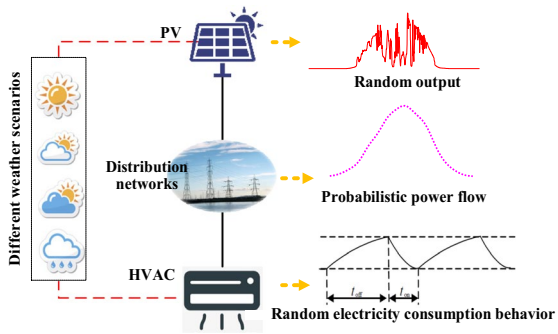

Fig. 1 Bilateral uncertainty in distribution networks

3 Methodology

To simulate stochastic optimal planning, several SML-based simulation modules are presented, including the scenario model, PPF model and planning method, as shown in Fig. 2.

The probabilistic scenario model simulates uncertainty, and the simulated data sets can be obtained on both the generation and demand sides. The above simulated data sets are sent to the PPF model, which estimates power flow responses considering the speed-accuracy trade-off effect. The planning model transforms the PPF information into probabilistic information of objectives and constraints for the eventual stochastic programming.

3.1 Probabilistic scenario model

The application of SML to the PV and HVAC model can be divided into two steps. First, the PV and HVAC models are constructed using SML-based models instead of traditional circuit models. A weather probability model is then constructed as input to the PV

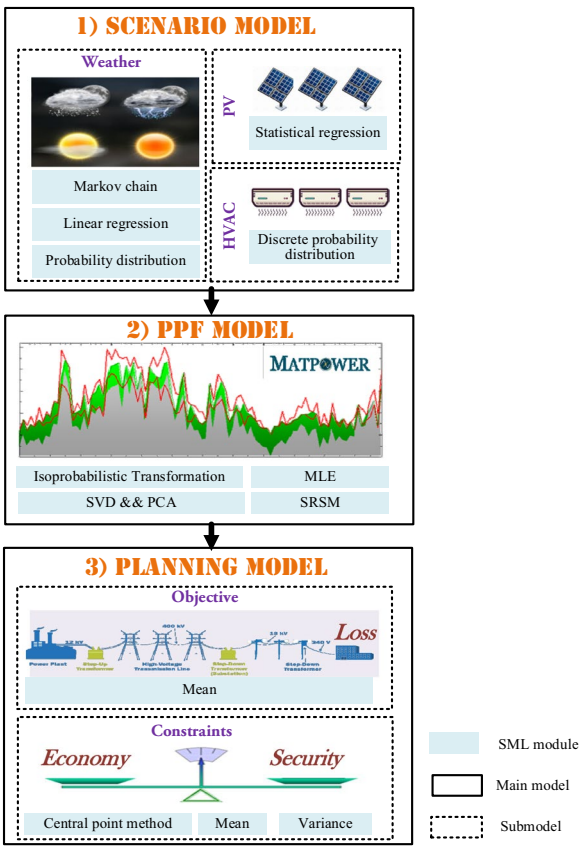

Fig. 2 SML-based simulation modules

and HVAC models. Because of the important impact of weather on PV power output and HVAC load, the existing proven models are based on weather rather than direct probabilistic modeling of power data [18, 19]. The advantage for such an approach is that strict physical constraints can be placed on the PV and HVAC models, and the generalizability can then be guaranteed. The PV and HVAC models for generating power samples are described in Table 1.

The PV model uses the solar radiation and outdoor temperature to calculate the PV power, while the HVAC model uses thermostat temperature setpoints and outdoor temperature to calculate the power loads. The distribution law of thermostat temperature setpoints represents HVAC electricity consumption behavior.

Remark 1 As the PV model is built on SML theory rather than a physical model, it is essentially a statistical regression model that is used to find the relationship between variables, i.e., PV power, solar radiation and temperature. The HVAC model depends on the distribution law of thermostat temperature setpoints, and the modeling method belongs to SML theories.

Table 1 PV and HVAC models

Ref	Model	Subject	SML
[17]	Discrete probability distribution	Consumption behavior	yes
[18]	Equivalent thermal model	HVAC	no
[19]	Linear regression	PV	yes

Table 2 Existing weather probability models

Ref	Mathematical theory	Subject
[20]	Correlation coefficient, Nataf transformation, Known probability distributions	Probability theory
[15]	Copula function, Known probability distributions	Probability theory
[16]	Copula function, Maximum entropy distributions	Probability theory and information theory
[17]	Copula function, Known probability distributions Markov chain	Probability theory and stochastic processes
Proposed	Nonlinear regression, Linear regression, Probability distribution, Markov chain	SML

To explain the concept of the proposed method clearly, the existing weather probability models are compared in Table 2.

Given the difference between the characteristics of the solar radiation and temperature curves, different statistical machine learning theories are proposed.

The temperature model is introduced first, where the hourly temperature series is modeled as a sum of two components, i.e., a deterministic component that explains the seasonal temperature and a stochastic component that explains predictive deviations. The deterministic component is modeled using nonlinear regression, i.e., a sum of sines, which represent the physical nature of the periodicity of temperature. A fit object is created to encapsulate the results of fitting the model specified by the sum of sine functions to the serial data, as:

$$T_{\text{fit}} = \sum_{i=1}^{n} a_i \times sin(b_i \times x + c_i), \tag{1}$$

where T_{fit} is a fit curve of temperature for a given hour in a given year, x is a vector of hourly dates, which is converted into serial date numbers. a_i is the amplitude, b_i is the frequency, c_i is the phase constant, and $n=2$ is the number.

The parameters of (1) are obtained using nonlinear least-squares. From (1), it follows that:

$$T_{\text{res}} = T_{\text{raw}} - T_{\text{fit}}, \tag{2}$$

where T_{res} is the residual, i.e., the stochastic component, while T_{raw} is the observed data for a given hour in a given year.

The stochastic component is modeled with a seasonal autoregressive model with seasonal lags, such that:

$$T_{\text{res},k} = a_0 + a_1 T_{\text{res},k-1} + \cdots + a_p T_{\text{res},k-p} + \varepsilon_k, \tag{3}$$

where ε_k is the white noise, and a_0, a_1, ..., a_p are the regression coefficients. The coefficients of the multiple linear regression are solved using least squares.

In a linear model, observed values are random variables, as are their residuals. Residuals have a t-location-scale distribution, which can be shown to provide a good fit, as:

$$PDF(T_{\text{res2}}) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \times \left[\frac{\nu + \left(\frac{T_{\text{res2}}-\mu}{\sigma}\right)^2}{\nu}\right]^{-\frac{\nu+1}{2}} \tag{4}$$

where $PDF(\cdot)$ is a probability density function, T_{res2} is the residual of (3), $\Gamma(\cdot)$ is the gamma function, μ is the location parameter, σ is the scale parameter, and ν is the

shape parameter. These are estimated using maximum likelihood estimates.

Following the temperature model, the solar radiation model is then described. The hourly temperature sample is modeled using a beta distribution [21]:

$$PDF(G) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{G}{G_{\max}}\right)^{\alpha-1}\left(1-\frac{G}{G_{\max}}\right)^{\beta-1}, \tag{5}$$

where α and β are the shape parameters, and G and $G_{\max}$ are the current and maximum solar radiations, respectively.

It is impossible to simulate seasonal characteristics and stochastic processes with only one probability distribution. Thus, the solar radiation model is improved using the Markov chain:

$$P_{ij}^{m} = P\{X_{n+k} = j | X_k = i\}, \quad m \geq 0,\ i, j \geq 0, \tag{6}$$

where p_{ij} is the one-step transition probability, $p^{m}{}_{ij}$ is the m-step transition probability, and the state is defined by splitting the beta CDFs of temperature samples. The Chapman-Kolmogorov equations provide a method for computing $p^{m}{}_{ij}$, as:

$$P_{ij}^{m+h} = \sum_{k=0}^{\infty} P_{ik}^{m} P_{kj}^{h} \ \text{ for all } m, h \geq 0, \ \text{ all } i, j, \tag{7}$$

The method of stochastic simulation of full-year solar radiation is as follows.

Step 1 Seasonality modeling.

- Divide the collected solar radiation in a given year into multiple seasonal intervals.
- Record the number where the solar radiation is not zero.

Step 2 Probability distribution estimation.

- Estimate the CDFs of the nonzero solar radiation using (5) for each seasonal interval.

Step 3 Seasonality modeling.

- Split the CDFs into several partitions for each seasonal interval.
- Compute the Markov chain empirical probability of going to state (j) from state (i) via statistical CDFs.
- Estimate empirical discrete distributions for each interval on each state.
- Create sample state path from empirical probability.
- Simulate a CDF when the above empirical discrete distribution is used in a simulated state.

Step 4 Simulate solar radiation throughout the year.

- Generate solar radiation using the inverse of the beta CDF for each seasonal interval.
- Replace the real nonzero solar radiation using the above simulated solar radiation.
- Connect the multi-segment seasonal simulation data in sequence.

Remark 2 PDF and CDF in probability theory are classical methods of uncertainty modeling for PPF calculation, but they become invalid for seasonal and dynamic characteristics. For modern active distribution networks, weather models can be simulated using an SML-based model, which helps improve the simulation of HVAC and PV.

3.2 PPF model

A high proportion of new energies and flexible demand-side resources make power flow uncertain, but it can be effectively analyzed using PPF. It is clear from numerous studies that PPF modeling does not need to consider the dynamic characteristics in either power generation or load demand. When PPF modeling depends on the power and load data calculated from the simulated weather data, the dynamic characteristics of weather data should be considered.

A series of SML algorithms, listed in Table 3, are adopted to calculate the PPF using the data of nondeterministic demand loads (HVAC loads) and DERs (PV power). The involved SML algorithms include principal component analysis (PCA), isoprobabilistic transformation, maximum likelihood estimator (MLE), singular value decomposition (SVD) and the stochastic response surface method (SRSM).

First, the isoprobabilistic transformation is adopted to transform the non-normal variables into standard normal variables, as:

$$p_i = F_i^{-1}(\Phi(x_i)), \tag{8}$$

where p_i is the PV power and HVAC load, and $F^{-1}(\cdot)$ is the estimated inverse cumulative distribution function. $\Phi(\cdot)$ is the standard normal CDF, and x is the standard normal variable. When the dimensionality of ν is not sufficiently low, the dimensionality of $\boldsymbol{X}$ should be reduced.

Second, the calculation formula of intrinsic dimensionality based on MLE is given by [22]:

$$\hat{d}_{MLE} = \frac{1}{k_2 - k_1 + 1}\sum_{k=k_1}^{k_2} \overline{d}_k, \tag{9}$$

Table 3 Probability distributions for PPF estimation

SML	Problem solved	Application description
Isoprobabilistic transformation	Standard normal	Step 1: Convert PV and HVAC data into standard normal data
MLE	Intrinsic dimensionality	Step 2: Calculate intrinsic dimensionality of normal data
SVD && PCA	Dimensionality reduction	Step 3: Reduce dimension of normal data to intrinsic dimensionality
SRSM	PPF estimation	Step 4: Estimate PPF using low dimensional data

$$\overline{d}_k = \frac{1}{n}\sum_{i=1}^{n}\hat{d}_k(x_i), \tag{10}$$

$$\overline{d}_k(x) = \left[\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(x)}{T_j(x)}\right]^{-1}, \tag{11}$$

where $\hat{d}_{MLE}$ is the intrinsic dimensionality of $\boldsymbol{X}$, k_1 *is* equal to 10, k_2 *is* equal to 20, and n is the sample size of V. $\overline{d}_k(\cdot)$ is the maximum likelihood estimator of the intrinsic dimensionality, and $T_j(\cdot)$ is the Euclidean distance from point x to the jth nearest neighbor within the hypersphere centered at x.

The number of DPFs, which are inputs and outputs of the SRSM, is equal to the sample size of p_i, which depends on the intrinsic dimensionality, as:

$$n_a = \frac{(n+p)!}{n!p!}, \tag{12}$$

$$l = 2 \times n_a, \tag{13}$$

where p is the order of the SRSM and l is the sample size of the uncertainty variables. It generates l points between 1 and n using:

$$ind = \text{linspace}(1, n, l) \tag{14}$$

where linespace $(\cdot)$ is a function that linearly generates spaced vectors, and *ind* is the serial number of scenarios whose size is equal to 8760. The sample size is reduced to $\boldsymbol{X}$ by retaining the sequence number *ind*. In addition, $\boldsymbol{X}$ becomes $\boldsymbol{X_1}$ by reducing the sample size. An important rule here is that the intrinsic dimension will affect the SRSM sample size, which is the same as the number of DPF calculations.

Third, a novel dimensionality reduction method is introduced. SVD produces a diagonal matrix $\boldsymbol{S}$ of the same dimension as X so that:

$$\mathbf{C} = \mathbf{USV}^{\mathrm{T}}, \tag{15}$$

$$\mathbf{C} = \text{cov}(\mathbf{X}_1), \tag{16}$$

where $\boldsymbol{C}$ is the covariance matrix of X, cov$(\cdot)$ is the covariance function, and the covariance matrix $\boldsymbol{U}$ is a l $\times$ l matrix. $\boldsymbol{S}$ is a diagonal matrix with l rows and n columns, and $\boldsymbol{V}$ has dimensions of $l \times l$. Note that $\boldsymbol{V}$ is the inverse of the square matrix $\boldsymbol{U}$.

From the above equations, the constructed matrix $\boldsymbol{Z}$ is obtained:

$$\mathbf{Z} = \mathbf{V}^{\mathbf{T}} \times (\mathbf{X}_1 - \boldsymbol{\mu}_{\mathbf{X}}), \tag{17}$$

where $\boldsymbol{\mu_X}$ stands for the mean of $\boldsymbol{X_1}$, the size of $\boldsymbol{\mu_X}$ is equal to the size of X_1, and $\boldsymbol{Z}$ is the constructed independent random variable.

The importance coefficient can be calculated using:

$$\gamma_i = \frac{s_i}{\sum_{i=1}^{m} s_i} \tag{18}$$

where m is the number of uncertainty variables, i.e., the dimensionality of the PV power and HVAC load. The dimensionality of $\boldsymbol{Z}$ is reduced by retaining the $\hat{d}_{MLE}$ dimensional importance uncertainty variables based on PCA theory. In addition, $\boldsymbol{Z}$ becomes $\boldsymbol{X_2}$ by reducing dimensionality, and an important rule here is that SVD and PCA reduce not only sample size but also dimensionality.

$\boldsymbol{X_2}$ should be standardized using:

$$\boldsymbol{\xi} = [\mathbf{X_2} - E(\mathbf{X_2})]/\sqrt{D(\mathbf{X_2})}, \tag{19}$$

where $E(\cdot)$ is the mean function, $D(\cdot)$ is the standard deviation function, and $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{\hat{d}_{MLE}}$ is the input of the SRSM.

Remark 3 SVD can help realize the decoupling of random variables, so independent random variables can be obtained. The independence is the usage premise of PCA for dimensionality reduction.

Fourth, a second-order SRSM is considered to compute PPF [23]:

$$E(y_i) = a_0, \tag{20}$$

$$V(y_i) = \sum_{i=1}^{K} a_i^2 + 2\sum_{i=1}^{K} a_{ii}^2 + \sum_{i=1}^{K-1}\sum_{j>i}^{K} a_{ij}^2, \tag{21}$$

where a_i is an unknown deterministic coefficient of SRSM, $V(\cdot)$ is the variance function, and y_i is a certain power flow response.

3.3 Planning model

A stochastic programming model is proposed as follows:

$$\min f_{\text{obj}}(S, num) = E(p_{loss})$$
$$s.t. \begin{cases} 0 \le S \le S^{-} \\ 2 \le num < num_{sys}, \\ \Pr(v_i > v_{-}) \ge \alpha \end{cases} \tag{22}$$

where $f_{\text{obj}}(\cdot)$ is the objective function, p_{loss} is the power loss, S is the capacitor capacity, and S^{-} is the upper boundary of the reactive capacity. num denotes the capacitor bus number, and num_{sys} is the bus number of the power network. v_i is the ith bus voltage amplitude, $\underline{v}$ is the voltage lower boundary, $\Pr(\cdot)$ is a probability function, and α is the confidence level.

In this planning model, power system operation is balanced, and power flow limits are considered in the DPF calculation. The formulas of power flow constraints are not given here. Only the formulas of bus voltage amplitude constraints are shown.

Equation (20) is used to directly calculate $f_{\text{obj}}(S,num)$, while (20) and (21) are used to calculate $\Pr(v_i > \underline{v})$ via a center point method, which is an SML. The limit state function is:

$$g(v_i) = v_i - v_{-}, \tag{23}$$

Equation (23) is expanded into a Taylor series at the center point, and the first-order term is retained as:

$$g(v_i) \approx g(\overline{v}_i) + (v_i - \overline{v}_i)^T \nabla g(\overline{v}_i), \tag{24}$$

$$\overline{g}(v_i) \approx g(\overline{v}_i), \tag{25}$$

$$\sigma\left(g(v_i)\right) \approx \sqrt{[\nabla g(\overline{v}_i)]^T C(\overline{v}_i) \nabla g(\overline{v}_i)}, \tag{26}$$

where $\overline{v}_i = \mathrm{E}(v_i)$, $C(\cdot)$ is the covariance function, $\overline{g}(v_i)$ is the mean value of $g(v_i)$, and $\sigma(\cdot)$ is the variance function. The structural reliability β is obtained by dividing (25) by (26):

$$\beta = \frac{g(\overline{v}_i)}{\sqrt{[\nabla g(\overline{v}_i)]^T C(\overline{v}_i) \nabla g(\overline{v}_i)}}, \tag{27}$$

$$\Pr\left(v_i > v_{-}\right) = 1 - \Phi(-\beta), \tag{28}$$

where $\Phi(\cdot)$ is the normal CDF.

4 Simulation

First, the simulation results of capricious weather models are introduced. These are inputs for PV and HVAC models. PPF is then estimated using different methods, and stochastic optimal planning of distribution networks is simulated.

4.1 Capricious weather simulation

In the simulation, annual weather data in 2015, including temperature and the solar radiation of Beijing, is collected. The data comes from [16], and the principle of maximum entropy (POME) distribution in [16] and normal distribution in [17] are used to verify the proposed temperature model. To verify the proposed solar irradiance model, a beta distribution in [21] is introduced as a reference. The sample size of simulation is 8760. The characteristics of the time series are analyzed first and then followed by the probability characteristics of the simulation results. As the probability model obviously cannot reflect the time series characteristics, it is no longer necessary to analyze the time series characteristics of the POME distribution in [16] and the normal distribution in [17].

As shown in Figs. 3 and 4, the proposed temperature model can simulate the stochastic time series correctly. The sample autocorrelation functions (ACFs) of the temperatures show that the time series properties and characteristics are well simulated. Descriptive statistics such as CDF, mean and standard deviation are introduced to test the probabilistic digital characteristics of temperature simulation models. As shown in Fig. 5 and Table 4, the normal distribution in [17] is calculated, and the proposed model is accurate according to the evaluation criterion of probability digital characteristics. The POME distribution in [16] can realize the judgment of the whole situation under the condition of missing data information, i.e., the CDF can be obtained by using moments. The above reasons lead to the difference between the CDF of the POME distribution and the empirical distribution function of the actual data. Lack of data information can necessitate a POME theory.

As shown in Fig. 6 and Table 5, it can be seen that the simulation results of the proposed model are accurate according to the evaluation criteria of dynamic characteristics and probability digital characteristics. The proposed model harnesses a Markov chain to obtain the dynamic characteristics of solar radiation fluctuations, while the CDF data generated by the Markov chain can grasp the probability characteristics of solar radiation. In addition, reasonable division of the whole year can ensure the seasonality of the solar radiation

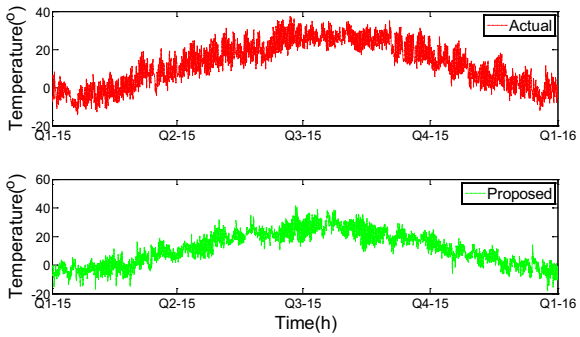

Fig. 3 Temperature values simulated based on the proposed model

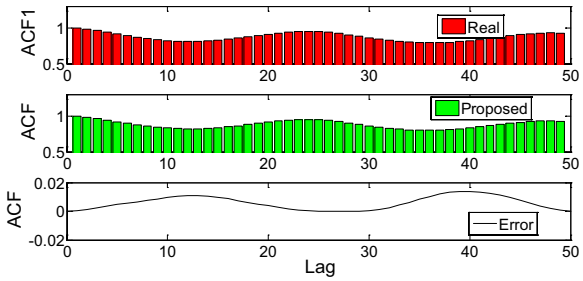

Fig. 4 Sample autocorrelation functions

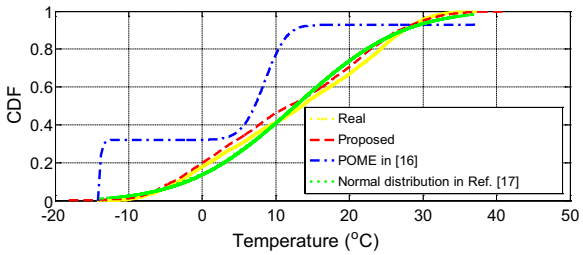

Fig. 5 CDFs of temperature simulated

Table 4 Descriptive statistics for simulated temperatures

Model	Mean/°C	Standard deviation/°C
Real	12.65	11.49
Proposed	11.76	11.68
Ref. [16]	15.58	10.64
Ref. [17]	12.71	11.29

data. The research here does not try to demonstrate that the previous uncertainty models are not good. Rather, it shows that the novel SML weather model is more effective in solving PPF problems.

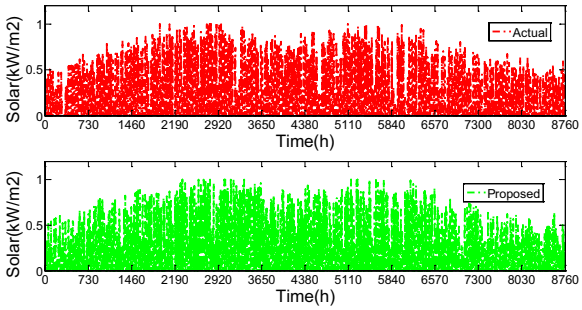

Fig. 6 Solar radiations simulated based on the proposed model

Table 5 Descriptive statistics for simulated temperatures

Model	Mean (kW/m^2)	Standard deviation (kW/m^2)
Real	0.16	0.24
Proposed	0.16	0.24
Ref. [17]	0.17	0.25

Discussion 1 The traditional methods in [15–17] do not take into account the seasonal variation of the scenarios, and thus lose the same data dependence performance in the month and season dimensions. The essence of PPF is to estimate the probability characteristics of state variables. It may be considered that it is sufficient to estimate PPF by mastering the probability characteristics of capricious weather variables. However, it is necessary to include the dynamic changes in the weather for HVAC loads due to building thermal inertia, since not only the current temperature but also past temperatures affect the HVAC loads. The stochastic process model for PV generation can improve the PPF calculation results of distribution networks with inertia HVAC loads, while the stochastic process of weather conditions should also be considered at the same time. The proposed SML can model both the stochastic process and probability characteristics.

4.2 PPF estimation simulation

The simulation data and parameters are elaborated as follows: (a) The simulated weather data in Section A are the input to the HVAC and PV models; (b) It is assumed that one 2 MW PV generation is installed at bus 3 in case 33 bw from MATPOWER, and the load of each PQ node is set to 1 kW base load plus 10 HVAC loads; (c) Each HVAC cools 140 areas, and other HVAC parameters are set according to those in [20]. The

Table 6 Probability law for customer thermostat setpoints

Setpoint (°C)	Probability	Setpoint (°C)	Probability	Setpoint (°C)	Probability
16	0.01	21	0.025	26	0.2
17	0.01	22	0.025	27	0.025
18	0.01	23	0.2	28	0.025
19	0.01	24	0.2	29	0.025
20	0.01	25	0.2	30	0.025

Table 7 Time consumption of the algorithms

Method	Scenario reduction (s)	SRSM (s)	PEM (s)	DPF (s)	Total (s)
FSA	0.00	0.00	0.00	106.21	106.21
PEM	0.00	0.00	0.22	1.16	1.39
Proposed	1.54	0.05	0.00	0.24	1.83

probability law of customer thermostat setting is listed in Table 6.

In this section, MATPOWER is selected to calculate DPF in a real scenario. The probability statistics of DPFs under 8760 scenarios can be called the full scenario approach (FSA), and the results of the FSA can be regarded as the correct results. In addition to the FSA, the point estimate method (PEM) in [23] is also compared with the proposed method. Time consumptions of the algorithms are listed in Table 7, and the listed CPU times can be explained by the number of scenarios and nondeterministic variables listed in Table 8.

Discussion 2 The calculation time of the PEM is determinable. The DPF number for the proposed method is 3.3 times that for the PEM. Note that more scenarios represent more DPF calculations and cost more CPU time, while the number of nondeterministic variables will not affect DPF calculation and its time. The calculation time of the proposed method depends on the nondeterministic variable intrinsic dimension, which determines the number of scenarios. These are the key explanations for the time consumption of the two methods.

For the analysis of calculation accuracy, the results of active power loss and voltage amplitude are provided. The means and standard deviations of the PQ bus voltages are shown in Figs. 7 and 8, respectively, while the loss means are listed in Table 9. As can be seen, the accuracy of the proposed method is similar to the PEM, with the proposed method being more accurate for some statistical indices while the PEM is more accurate for other statistical indices. As shown in Fig. 9, the curve of the proposed method is close to the curve of the real results, while the curve of the PEM in * MERGEFORMAT [24] is significantly biased.

A realistic power distribution system named Jiaokeng in Guangdong, China from [15] is included to verify the proposed method, as shown in Fig. 10. The voltage of bus 0 is 10.5 kV, and the parameters of the PV and HVAC remain unchanged, while the samples are re-simulated. The means and standard deviations of the PQ bus voltages are shown in Figs. 11 and 12, respectively, while the

Table 8 Number of scenarios and nondeterministic variables

Method	Number of scenarios	Number of nondeterministic variables
FSA	8760	33
PEM	66	33
Proposed	20	3

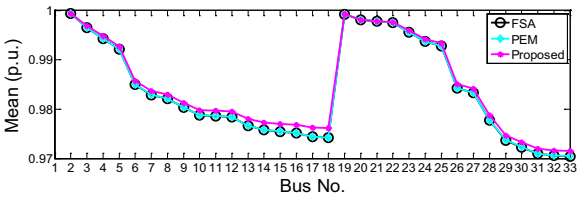

Fig. 7 Mean value of the elements in bus voltage profiles

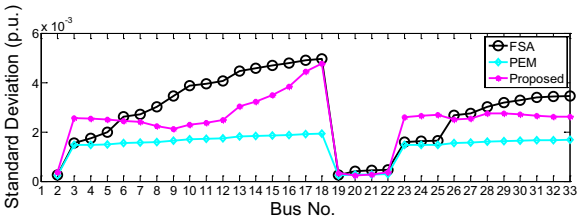

Fig. 8 Standard deviation of the elements in bus voltage profiles

means of the power loss are listed in Table 10. It can be concluded from the results that the proposed method can be used in actual distribution networks.

Discussion 3 By comparing with the PEM method, the values of the proposed method can be summarized. (a) The usability of PEM depends on the correctness of the DPF, whose parameter errors can lead to errors in the PPF results. In contrast, the usability of the proposed method depends on the DPF data rather than the DPF model. As in China's distribution networks, impedance parameters have not been correctly verified, and thus, the proposed method is needed. (b) The proposed method

Table 9 Means and standard deviations of loss (MW)

Method	Mean
FSA	0.0619
PEM	0.0614
Proposed	0.0608

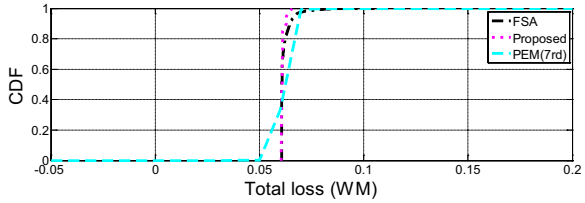

Fig. 9 CDFs of the total losses for different methods

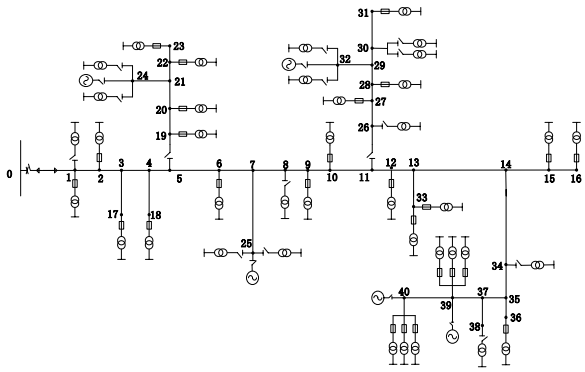

Fig. 10 Single-line diagram of the real power system

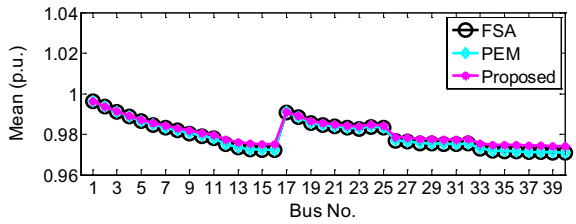

Fig. 11 Mean value of the PQ bus voltages

has an advantage over the PEM in estimating the CDFs of power flow responses, since the CDF information of the PEM is from moments while for the proposed method, it is from power flow responses. (c) Both the PEM and proposed methods cannot exactly match the real results, while the extraction of key information based on SML also results in information loss. However, as the accuracy is guaranteed and the efficiency is greatly improved, the proposed method has application value and is consistent with the idea that machine learning should balance robustness and bias.

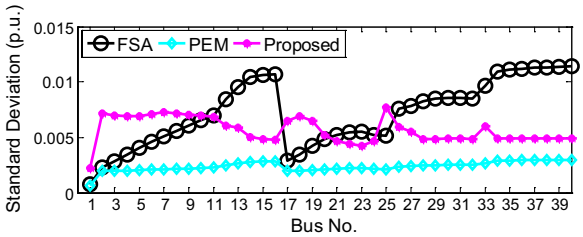

Fig. 12 Standard deviation of the PQ bus voltages

Table 10 Means of the power loss (MW)

Method	Mean
FSA	0.022
PEM	0.020
Proposed	0.018

4.3 Stochastic programming simulation

This part of the simulation demonstrates the PPF-based power planning solution, to show the practical engineering significance of PPF calculation. It verifies the conservatism of inequality probability inequality (PI) theory in [17]. Being too conservative will lead to an insufficient economy, which is the motivation of this paper. To verify probability inequality, 8760 DPFs are calculated for the whole year, and the simulation results are listed in Table 11.

The fundamental purpose of the proposed method is to reduce the computing time of objective and constraint functions for each group of solutions. The essence of an efficient planning model is that its calculation efficiency is greatly improved under the premise of small calculation accuracy loss. A given planning scheme is used to verify the proposed method, as shown in Table 12 and Fig. 13.

Note that the SML does not blindly pursue the small deviation but also balances the deviation and generalizability. The proposed PPF method reduces 8760 scenarios to 20, and the central point method uses only mathematical expectations and variances. After these two steps, the calculation efficiency has been greatly improved. Although the calculation accuracy is slightly reduced, the feasibility and efficiency of the planning solution are guaranteed. Conservatism of the central point method is better than probability inequality and can be used in planning. The PSO algorithm in [17] is used to solve the proposed programming model, and the simulation results are listed in Table 13. As can be seen, compared with the total loss in Table 9, the optimum total loss in Table 13 is much smaller.

Table 11 Simulation results using theory in [17]

Item	Value
Planning location	Buses 7 and 8
Planning capacity	[100 kVAr, 100 kVAr]
Voltage qualification limits	[0.9 p.u., 1.1 p.u.]
Worst bus probability for FSA	1
Worst bus qualification probability for PI	0.9223

Table 12 Simulation results using the proposed method

Item	Value
Planning location	Buses 7 and 8
Planning capacity	[100 kVAr, 100 kVAr]
Lower voltage limit	0.95 p.u
Cost time for FSA	220.2490 s
Cost time for the proposed method	0.6190 s
Objective function for FSA	0.0525 MW,
Objective function for the proposed method	0.0513 MW

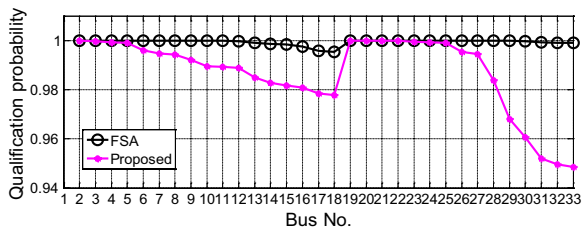

Fig. 13 Bus voltage qualification probabilities for different methods

Table 13 Optimum planning results

Item	Value
Reactive capacity boundary	0–5000 kVAr
Number of particles	20
Number of iterations	10
Optimum planning location	Buses 8 and 30
Optimum planning capacity	[478 kVAr, 935 kVAr]
Objective function	0.0066 MW
Lower voltage limit	0.95 p.u
Worst bus confidence level	0.9984
Planning solution time	260 s

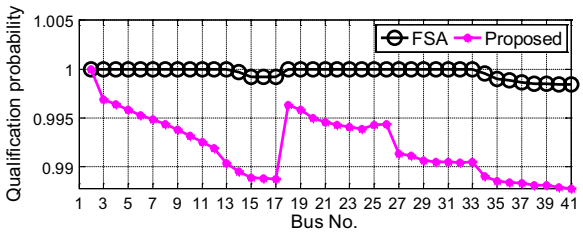

Fig. 14 Bus voltage qualification probabilities

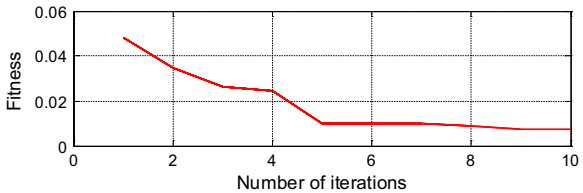

Fig. 15 Fitness value of global optimal solution via iteration times

A real 41-bus distribution network in Guangdong, China is included to verify the proposed method. The planning scheme remains unchanged except for changing the voltage lower limit to 0.9 p.u. Bus voltage qualification probabilities are shown in Fig. 14, the fitness value of global optimal solution is shown in Fig. 15, and the planning results are listed in Table 14. As shown, compared with the total loss in Table 14, the optimum total loss in Table 10 is much smaller. Thus, the proposed method can be used for planning actual distribution networks.

The innovation of this paper is to establish an efficient planning model rather than PSO (i.e., a mathematical programming solver), and thus, its contribution is to improve the accuracy and efficiency of calculating a feasible solution via improving the planning model rather than improving a mathematical programming algorithm.

Discussion 4 Smart grid planning has economic and technical indicators. In terms of economic indicators (such as network loss), mathematical expectation can be used as an effective measure. For the technical index (such as voltage deviation), the boundary condition of the index rather than the probability characteristic is a problem of concern. Thus, PPF calculation results can be used directly for economic indicators, but it becomes challenging to apply them to technical indicators. Although it is feasible to transform the PPF probabilistic information into boundary information, the conservatism of the transformation results is natural regardless of the adopted mathematical theory. It is reliable to apply PPF to stochastic programming via the central point method, which limits the probability of unqualified voltage deviation to a certain range.

Table 14 Planning results

Item	Value
Optimum planning location	Buses 14 and 21
Optimum planning capacity	[337 kVAr, 398 kVAr]
Objective function	0.007 MW
Worst bus confidence level	0.9966
Planning solution time	301 s

5 Discussion

The uncertainty of renewable energy can cause problems in the operation and planning of electric distribution networks. Existing literature has highlighted probability theory in dealing with such uncertainty, and recent research demonstrates that a probability model can deal with the uncertainty of the distribution networks well when there are only small numbers of renewable energy plants. However, it becomes very complex, time-consuming and error-prone to develop and infer the stochastic planning model of distribution networks based on statistics. With large numbers of renewable energy plants, modeling uncertainty becomes complex and cannot be handled by traditional methods. SML is oriented to algorithms and attaches importance to prediction results. From the study, it can be concluded that the SML-based planning model has good controllability and scalability, and can overcome the limitations of the traditional statistical model development and inference algorithms in distribution network stochastic planning, thus realizing the in-depth development of SML in the field of renewable energy integration.

6 Conclusions

A distribution network with large penetration of new energy is a large-scale high-dimensional dynamic system with nonlinear, uncertain and complex characteristics. High-dimensional nonlinearity and uncertainty bring difficulties and challenges to the refined analysis of operating performance and optimal planning solutions in distribution networks. In this paper, statistical machine learning techniques are introduced to carry out multi-scenario based probabilistic power flow calculations and are applied to the stochastic planning of distribution networks. An SML-based capricious weather model is established to improve accuracy, and a series of techniques are adopted to promote the efficiency of PPF estimation with the PPF probabilistic information transformed into boundary information for the eventual stochastic programming. Both the IEEE 33-bus system and a real distribution network are studied to validate the proposed method. Simulation results show that the proposed SML-based planning model performs better than traditional statistical models and algorithms in distribution network stochastic planning. Thus, the SML-based planning is adequate and has the potential for practical application.

Acknowledgements
The author would like to thank the referees and the editor of this journal for valuable comments.

Authors' contributions
XF designed research, performed research, analyzed data, and wrote the paper. The author read and approved the final manuscript.

Authors' information
Xueqian Fu received his B.S. and M.S. degrees from North China Electric Power University in 2008 and 2011, respectively. He received his Ph.D. degree from South China University of Technology in 2015. From 2011 to 2015, he was an electrical engineer with Guangzhou Power Supply Co. Ltd. From 2015 to 2017, he was a Post-Doctoral Researcher with Tsinghua University. He is currently an Associate Professor with China Agricultural University. His current research interests include Agricultural Energy Internet and Statistical Machine Learning. He is currently an Editor for Power Demand Side Management. He chaired a session in 2020 Asia Energy and Electrical Engineering Symposium (IEEE-AEEES 2020).

Funding
This study is supported by the National Natural Science Foundation of China under Grant 52007193 and The 2115 Talent Development Program of China Agricultural University.

Availability of data and materials
The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 9 December 2021 Accepted: 7 February 2022
Published online: 18 February 2022

References

1. 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.
2. Liu, S., Zhou, C., Guo, H., et al. (2021). Operational optimization of a building-level integrated energy system considering additional potential benefits of energy storage. *Protection and Control of Modern Power Systems, 6*, 4.
3. Zhang, C., Chen, H., Shi, K., Qiu, M., Hua, D., & Ngan, H. (2018). An interval power flow analysis through optimizing-scenarios method. *IEEE Transactions on Smart Grid, 9*(5), 5217–5226.
4. Minchala-Avila, L. I., Garza-Castañon, L., Zhang, Y., & Ferrer, H. J. A. (2016). Optimal energy management for stable operation of an islanded microgrid. *IEEE Transactions on Industrial Informatics, 12*(4), 1361–1370.
5. Yu, J., Dai, W., Li, W., Liu, X., & Liu, J. (2018). Optimal reactive power flow of interconnected power system based on static equivalent method using border PMU measurements. *IEEE Transactions on Power Systems, 33*(1), 421–429.

6. Dai, W., Yu, J., Yang, Z., Huang, H., Lin, W., & Li, W. (2020). A static equivalent model of natural gas network for electricity–gas co-optimization. *IEEE Transactions on Sustainable Energy, 11*(3), 1473–1482.
7. Zhang, H., Hu, Z., Xu, Z., & Song, Y. (2017). Optimal planning of PEV charging station with single output multiple cables charging spots. *IEEE Transactions on Smart Grid, 8*(5), 2119–2128.
8. Chen, J., Xiao, Y., Mo, R., & Tian, Y. (2021). Optimized allocation of microgrid energy storage capacity considering photovoltaic correction. *Power System Protection and Control, 49*(10), 59–66.
9. Yan, C., Tang, Y., Dai, J., et al. (2021). Uncertainty modeling of wind power frequency regulation potential considering distributed characteristics of forecast errors. *Protection and Control of Modern Power Systems, 6*, 22.
10. Wang, J., Zhong, H., Xia, Q., & Kang, C. (2018). Optimal planning strategy for distributed energy resources considering structural transmission cost allocation. *IEEE Transactions on Smart Grid, 9*(5), 5236–5248.
11. Hamad, A. A., Nassar, M. E., El-Saadany, E. F., & Salama, M. M. A. (2019). Optimal configuration of isolated hybrid AC/DC microgrids. *IEEE Transactions on Smart Grid, 10*(3), 2789–2798.
12. Zhang, C., Li, J., Zhang, Y. A., & Xu, Z. (2020). Data-driven sizing planning of renewable distributed generation in distribution networks with optimality guarantee. *IEEE Transactions on Sustainable Energy, 11*(3), 2003–2014.
13. Fu, X., Chen, H., Cai, R., & Yang, P. (2015). Optimal allocation and adaptive VAR control of PV-DG in distribution networks. *Applied Energy, 137*, 173–182.
14. Fu, X., Chen, H., Xuan, P., & Cai, R. (2016). Improved LSF method for loss estimation and its application in DG allocation. *IET Generation, Transmission & Distribution, 10*(10), 2512–2519.
15. Fu, X., Sun, H., Guo, Q., Pan, Z., Zhang, X., & Zeng, S. (2017). Probabilistic power flow analysis considering the dependence between power and heat. *Applied Energy, 191*, 582–592.
16. Fu, X., Sun, H., Guo, Q., Pan, Z., Xiong, W., & Wang, L. (2017). Uncertainty analysis of an integrated energy system based on information theory. *Energy, 122*, 649–662.
17. 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.
18. Lu, N. (2012). An evaluation of the HVAC load potential for providing load balancing service. *IEEE Transactions on Smart Grid, 3*(3), 1263–1270.
19. Rohani, G., & Nour, M. (2014). Techno-economical analysis of stand-alone hybrid renewable power system for Ras Musherib in United Arab Emirates. *Energy, 64*, 828–841.
20. Chen, Y., Wen, J. Y., & Cheng, S. J. (2013). "Probabilistic load flow method based on Nataf transformation and Latin hypercube sampling. *IEEE Transactions on Sustainable Energy, 4*(2), 294–301.
21. Karaki, S. H., Chedid, R. B., & Ramadan, R. (1999). Probabilistic performance assessment of autonomous solar-wind energy conversion systems. *IEEE Transactions on Energy Conversion, 14*(3), 766–772.
22. Levina, E., & Bickel, P.J. (2004). Maximum likelihood estimation of intrinsic dimension. In Advances in neural information processing systems, vol. 17. Cambridge: The MIT Press.
23. Chun-Lien, Su. (2005). Probabilistic load-flow computation using point estimate method. *IEEE Transactions on Power Systems, 20*(4), 1843–1851.
24. Fu, X., Wu, X., & Liu, N. (2021). Statistical machine learning model for uncertainty planning of distributed renewable energy sources in distribution networks. *Frontiers in Energy Research, 9*, 809254.