Hindawi
International Transactions on Electrical Energy Systems
Volume 2022, Article ID 9453879, 11 pages
https://doi.org/10.1155/2022/9453879

Research Article

Data-Driven Electricity Price Risk Assessment for Spot Market

En Lu,[1] **Ning Wang**,[1] **Wei Zheng**,[1] **Xuanding Wang**,[1] **Xingyu Lei**,[2] **Zhengchun Zhu**,[2] **and Zhaoyu Gong**[2]

[1]*Guangdong Electric Power Trading Center Co., Ltd., Guangzhou 510080, Guangdong Province, China*
[2]*Beijing Tsintergy Technology Co., Ltd., Haidian District, Beijing 100084, China*

Correspondence should be addressed to Xingyu Lei; lxylxy7@163.com

Received 7 October 2021; Accepted 23 November 2021; Published 31 January 2022

Academic Editor: Qiuye Sun

Electricity price risk assessment (EPRA) is essential for spot market analysis and operation. The statistical moments (i.e., the mean and standard deviation) of the price need to be assessed to support market risk control. This paper proposes a data-driven approach for EPRA based on the Gaussian process (GP) framework. Compared with the deep learning algorithms, GP has two merits: (1) the scale of training sample required is small and (2) the time-consuming hyperparameter tuning process is avoided. However, the direct application of GP for EPRA is not tractable due to the complicated discrete relationship between the system operating status and the electricity price. To deal with that, a data-driven EPRA framework is developed that contains a GP surrogate model for the direct current optimal power flow (DC-OPF) problem and a hybrid model-data-based hybrid electricity price calculation method. To guarantee the accuracy of EPRA, an adaptability criterion and a second verification process based on the Karush–Kuhn–Tucker (KKT) condition are developed to distinguish the samples with GP learning errors. Numerical results carried out on IEEE benchmark systems demonstrate that the proposed method can achieve exactly the same EPRA results as Monte Carlo (MC) simulation, which significantly improved the computational efficiency.

1. Introduction

To reduce pollution and greenhouse gas emissions, a high share of renewable energy integration has become one of the basic characteristics of the smart grid [1–3]. With the development of renewable energy and the adoption of locational marginal pricing (LMP) methodology, the spot market is full of uncertainties, such as load deviation and renewable variation [4]. The abovementioned uncertainties cause the electricity price to fluctuate violently, bringing significant operational and planning risks for electricity market participants.

Risk assessment can provide power system operators with prior knowledge and theoretical basis to ensure safe and stable power system operation [5–7]. In the spot market, electricity price risk assessment (EPRA) is crucial for independent system operators (ISOs) and market participants. However, it is more volatile and challenging to predict the fluctuations in electricity prices than the uncertainties of power production and consumption [8]. The reliable and efficient assessment method is the basis for spot market operation and risk control. Current studies focus on the risk caused by the electricity price fluctuation for the risk assessment in electricity markets. Reference [9] analyses the optimal electricity procurement problem for large consumers considering the electricity price fluctuation. Reference [10] proposes a value-at-risk (VaR) and conditional VaR (CVaR) assessment for electricity price risk based on historical data. Reference [4] uses the Monte Carlo simulation method in electricity price risk management. Reference [11] analyses the price risk of power portfolios in multimarkets based on the well-established mean-variance model.

For EPRA, the expectation and standard deviation of LMP need to be assessed to support market risk control of independent system operators (ISOs) [12]. Generally, LMP can be obtained based on the direct current optimal power flow (DC-OPF) model, which is derived from the Lagrangian multipliers of the power balance constraint and transmission constraints, including an energy component

and a congestion component [13]. Probabilistic optimal power flow (POPF) is able to comprehensively consider various uncertainties in the spot market and thus has become an effective tool to estimate LMP in the deregulated market [14, 15].

To solve the POPF problem, two main calculation approaches have been developed, namely, model-based and data-based approaches. The model-based methods can be roughly divided into analytical methods and simulation methods. Typical analytical methods, such as the point estimation method, construct representative samples according to the probability density function (PDF) of the uncertainty variables [16, 17]. EPRA results can be obtained according to the OPF solutions of representative samples, which is computationally efficient, but complicated mathematical derivations and strict assumptions are required. Typical simulation methods such as Monte Carlo (MC) simulations obtain EPRA results by using massive random generated samples that are carried out on the OPF model, which is reliable but computationally demanding [18, 19]. Recently, data-based machine learning methods have been widely applied in power system [20, 21], showing a promising way to achieve EPRA with high precision and fast computational speed. For POPF problem, the core idea of the data-based approach is to construct a data-driven surrogate model that treats the OPF problem as a functional mapping between the system operating status and the OPF solutions, thus greatly improving the computational efficiency of the POPF problem. In [22, 23], a deep neural network (DNN) approach for solving OPF problems was developed based on historical data and offline simulations. Reference [24] proposed a data-driven machine learning framework for the OPF problem considering the characteristics of the physical model. However, these data-driven approaches have several technical challenges between the POPF problem and EPRA. Several challenges need to be addressed. First, the discrete features of LMP are hard to be learned by the existing data-driven methods. Second, the data-driven methods, such as DNN-based approaches, usually require massive training samples, which may not align with the current spot market practice. Third, the inherent learning error of the data-driven methods is inevitable and may yield an unreliable EPRA result. To overcome the challenges mentioned above, our work combines the advantages of the two aforementioned approaches to develop a data-driven assisted electricity price risk assessment method based on the Gaussian process (GP) and the physical model of DC-OPF.

Compared with traditional DNN-based methods [25, 26], the GP is a novel machine learning technology that requires smaller training samples and fewer hyperparameters for learning [27, 28], making the GP align well with current industry practice. The GP is used extensively as a nonparametric regression tool in various scenarios, e.g., active learning [29], multitask learning [30, 31], manifold learning [32], and optimization [33]. However, the learning error of GP is also inevitable. Further advanced technology is required to accurately learn the features of LMP.

The objective of EPRA is to obtain the statistical moments of the LMP according to various uncertainties of the system operating status. Data-driven methods can build a surrogate model with cheap computation cost to replace the time-consuming LMP calculation process. Note that an efficient data-driven EPRA algorithm needs not only high precision and fast computing speed but also good generalization capability with limited training sample, which makes the unique characteristics of GP (e.g., fast training, less intervention, and small sample requirement) an ideal candidate. However, unlike the POPF problem, the relationship between the input (the system operating status) and the output (the LMP) is rather complex due to the discontinuous property of LMP. Hence, direct learning LMP using data-driven methods is intractable, which will be shown in the simulation results. Fortunately, the physical model of DC-OPF is known, and this motivates us to develop a new framework to achieve the LMP assessment based on the POPF results by including its physical characteristics.

To this end, a data-driven EPRA approach is proposed based on both physical models and historical data. Compared with existing methods, the proposed data-driven method combines the advantages of the model-based and data-based approaches to achieve a more efficient EPRA without accuracy loss. Specifically, we embed the GP surrogate model for DC-OPF into the model-based EPRA process to improve the computational efficiency of the traditional model-based method. By providing the strict judging criteria (adaptability criterion and a second verification) to determine the inaccurate samples obtained by the proposed data-driven method, the accuracy of EPRA is guaranteed. Note that the proposed approach is a general method for EPRA, even in a specific scenario with limited samples. It has the following advantages: (1) the accuracy of EPRA is maintained. The EPRA results obtained through our approach are exactly the same as those of the MC method. (2) The training sample size for learning the LMP has significantly reduced thanks to the GP. (3) The efficiency of EPRA is improved because a large proportion of the time-consuming POPF process is replaced by direct GP mapping.

The main contributions of this paper are summarized as follows:

(1) A data-driven framework is proposed to reduce the scale and accelerate the computational speed of the EPRA problem. Specifically, to avoid directly learning the LMP, a GP surrogate model for the DC-OPF problem is developed to identify key information for LMP calculation (e.g., the marginal generators and congested transmission lines). Then, a model-data hybrid EPRA method is proposed by solving a set of linear equations. The proposed method can significantly improve the efficiency of the EPRA without compromising its accuracy.

(2) Under this framework, a model-based adaptability criterion and a second verification for EPRA are developed to determine inaccurate samples. Before using the sample with marginal generators and congested transmission lines identified by the GP to

calculate the LMP, physical model information is used to distinguish the samples with learning errors. Hence, the accuracy of EPRA is guaranteed.

The rest of the paper is organized as follows: the data-driven EPRA framework is developed in Section 2. Section 3 presents the proposed GP surrogate model for DC-OPF. Numerical results are analyzed in Section 4, and finally, Section 5 concludes the paper.

2. The Data-Driven Framework for EPRA

In the spot market, the LMP arises from an economic dispatch. Specifically, the system operator solves a DC-OPF problem for the optimal economic generation that meets the variational load and renewable energy while satisfies the generation and transmission constraints [34]. In fact, there is a linear relationship between the LMP and the Lagrangian multiplier of the DC-OPF model. The relationship relies on the marginal generator and congested transmission line, which can be obtained through the DC-OPF solutions. Hence, the key idea of the proposed data-driven framework is to build a GP surrogate model for the DC-OPF problem to identify the marginal generator and congested transmission line, thus improving the computational efficiency of EPRA. The physical characteristics of the DC-OPF model are considered to ensure accuracy. In this section, the LMP formulation is first studied using a general DC-OPF formulation and its Karush–Kuhn–Tucker (KKT) condition. Then, the data-driven approach is proposed for EPRA.

Note that this paper is focused on the LMP risk arised from the uncertainty of load and renewable energy. Within the proposed scope, we assume the topology of power grid is invariable. The uncertainties from equipment fault are ignored.

2.1. A General DC-OPF Formulation. A general DC-OPF model for economic dispatch is formulated as follows:

Objective function:

$$\min \sum_{i\in I_G} c_i P_i. \tag{1}$$

Constraints:

$$\sum_{i\in I_G} P_i - \sum_{i\in I_D} D_i = 0 \colon \lambda, \tag{2}$$

$$\sum_{i\in I} PT\ DF_{li} \times \left(P_i - D_i\right) \le F_l \colon \eta_l^{\max}, \tag{3}$$

$$-\sum_{i\in I} PT\ DF_{li} \times \left(P_i - D_i\right) \le F_l \colon \eta_l^{\min}, \tag{4}$$

$$P_{\min,i} \le P_i \le P_{\max,i} \colon \xi_i^{\min}, \xi_i^{\max}, \tag{5}$$

where c_i is the generator cost for production; $PTDF_{li}$ represents the power transfer distribution factor of Bus i to Line l; $\sum_{i\in I} PT\ DF_{li} \times (P_i - D_i)$ is the transmission power flow of Line l, which is denoted as PF_l; and P_i and D_i are the generator output and demand quantity, respectively. Note that renewables are treated as negative loads in this paper, which are included in D_i.

The linear objective function of the DC-OPF model is designed to minimize the operating costs associated with supplying real power to meet the demand requirement. Equation (2) is the system power balance equation, and λ is the corresponding Lagrangian multiplier. The constraints in (3) and (4) limit the transmission line power flow, and $\eta_l^{\max}$ and $\eta_l^{\min}$ are the Lagrangian multipliers of the upper and lower transmission limit constraints, respectively. The constraints in (5) are the operational limits for the real generator power, and $\xi_i^{\max}$, $\xi_i^{\min}$ are the Lagrangian multipliers of the upper and lower limits of the generator output constraints, respectively.

2.2. Deduction for the LMP Formulation. To understand the internal relationship between the LMP and DC-OPF problem, the KKT condition is used to analyze the properties of LMP.

2.2.1. The LMP Formulation. According to the KKT condition, we derive the relationships among the LMP, the Lagrangian multiplier of the power balance λ, and the dual multiplier of the transmission line limits $\mu_l^{\max}$ and $\mu_l^{\min}$. Note that in the following analysis, the saddle point used by the KKT condition corresponds to the global optimum of the OPF model.

To obtain the LMP for the EPRA, the Lagrangian function of the DC-OPF models (1)–(5) is denoted by LF, as follows:

$$\begin{aligned} L = & \sum_{i\in I_G} c_i P_i - \lambda \sum_{i\in I} \left(P_i - D_i\right) \\ & - \sum_{l\in I_L} \eta_l^{\min} \left(\sum_{i\in I} PT\ DF_{li} \times \left(P_i - D_i\right) + F_l \right) \\ & - \sum_{l\in I_L} \eta_l^{\max} \left(-\sum_{i\in I} PT\ DF_{li} \times \left(P_i - D_i\right) + F_l \right) \\ & - \sum_{i\in I_G} \xi_i^{\min} \left(P_i - P_{\min,i}\right) - \sum_{i\in I_G} \xi_i^{\max} \left(P_{\max,i} - P_i\right). \end{aligned} \tag{6}$$

Then, the LMP for the load at Bus i is derived from the Lagrangian function in (6) as

$$LMP_i = \frac{\partial L}{\partial D_i} = \lambda + \sum_{l\in L} PT\ DF_{li} \left(\eta_l^{\min} - \eta_l^{\max}\right). \tag{7}$$

From (7), we note that the LMP is obtained by the marginal generator through λ and the congested transmission line through $\eta_l^{\max}$ and $\eta_l^{\min}$. The relationship between the marginal generators and congested transmission lines is discussed in the next section.

2.2.2. The Relationship between Marginal Generators and Congested Transmission Lines. Based on the KKT condition, the following equation for generator i can be obtained:

$$\frac{\partial L}{\partial P_i} = c_i - \lambda - \sum_{l \in I_L} PT\ DF_{li}\left(\eta_l^{\min} - \eta_l^{\max}\right) - \left(\xi_i^{\min} - \xi_i^{\max}\right) = 0. \quad (8)$$

For the marginal generators, $\xi_i^{\max} = \xi_i^{\min} = 0$, and equation (8) can be expressed as

$$\lambda + \sum_{l \in I_L} PT\ DF_{li}\eta_l = c_i, \quad i \in I_{MG}, \quad (9)$$

where $\eta_l = \eta_l^{\min} - \eta_l^{\max}$ and I_{MG} and N_{MG} are the set and the number of marginal generators, respectively. Then, the matrix form of the equations above is analyzed. The matrix form of (9) is

$$\begin{bmatrix} 1 & PT\ DF_{11} & \cdots & PT\ DF_{L1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & PT\ DF_{L1} & \ldots & PT\ DF_{LN_{MG}} \end{bmatrix}_{N_{MG} \times (L+1)} \times \begin{bmatrix} \lambda \\ \eta_1 \\ \vdots \\ \eta_L \end{bmatrix}_{(L+1)\times 1} = \begin{bmatrix} c_1 \\ \vdots \\ c_{N_{MG}} \end{bmatrix}_{N_{MG}\times 1}. \quad (10)$$

There are N_{MG} equations and $(L+1)$ unknown variables in (10), which cannot yield a unique solution. To solve these equations, $(L+1-N_{MG})$ variables should be determined in advance. Hence, the information of the transmission line constraints is introduced. Note that for transmission lines without congestion, $\eta_l = 0$.

Hence, the number of congested transmission lines N_{CL} is

$$N_{CL} = L - \left(L + 1 - N_{MG}\right) = N_{MG} - 1. \quad (11)$$

Then, the equations in (10) can be rewritten as

$$\begin{bmatrix} 1 & PT\ DF_{11} & \cdots & PT\ DF_{L1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & PT\ DF_{N_{CL}1} & \ldots & PT\ DF_{N_{CL}N_{MG}} \end{bmatrix}_{N_{MG} \times N_{MG}} \times \begin{bmatrix} \lambda \\ \eta_1 \\ \vdots \\ \eta_{N_{CL}} \end{bmatrix}_{N_{MG}\times 1} = \begin{bmatrix} c_1 \\ \vdots \\ c_{N_{MG}} \end{bmatrix}_{N_{MG}\times 1}. \quad (12)$$

Based on the above discussion, we know that if the marginal generators and the congested transmission lines are known, the LMP can be calculated by solving a set of linear equations in (12). Hence, the identification of marginal generators and congested transmission lines is the key problem for EPRA.

2.3. The Proposed Framework. In this section, a trustworthy data-driven framework is proposed for EPRA. Based on historical data, a GP surrogate model is developed for DC-OPF to identify the marginal generators and congested transmission lines, which will be introduced in the next section. After identification, the LMP can be obtained immediately without solving the time-consuming DC-OPF problem. Unfortunately, inherited learning error in data-driven methods, including the proposed GP surrogate model, is unavoidable, making the identification of marginal generators and congested transmission lines unreliable. To overcome this challenge, an adaptability criterion is developed based on the KKT condition of the physical model discussed in Section 2.2. The proposed framework for EPRA is illustrated in Figure 1. The three steps are described as follows:

Step 1: Marginal Generators and Congested Transmission Line Identification. The power system operating data (e.g., D_i) are collected from historical data or by running a Monte Carlo simulation, and they are input into the trained GP surrogate model for DC-OPF. For each sample, the DC-OPF outputs (e.g., the generator output and transmission power flow) are immediately obtained. Then, the marginal generators and congested transmission lines can be identified with high precision and fast computation.

Step 2: Distinguishing the Error Samples. According to the discussion presented in Section 2.2, for the DC-OPF problem, we note that the number of marginal generators N_{MG} is $n+1$ if the number of congested transmission lines N_{CL} is n. This natural property motivates us to adapt it to the proposed framework to distinguish the samples with learning errors. Hence, for each sample with marginal generators and congested transmission lines identified by the GP surrogate model, the proposed adaptability criterion is as follows:

$$N_{MG} + 1 = N_{CL}. \quad (13)$$

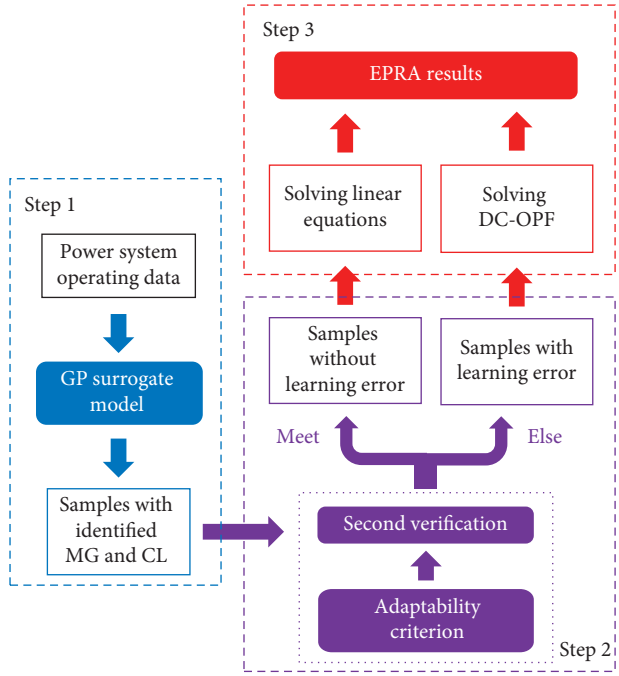

Figure 1: Flowchart of the proposed data-driven framework.

It should be noted that the proposed adaptability criterion is not strictly complete. There are a very few samples that meet the adaptability criterion but have learning errors. For these samples, the LMP error of all the buses can diverge significantly from the real value because the marginal cost of generation is changed. Hence, a second verification process is proposed based on historical data, as follows:

$$\frac{\left|LMP_i - \text{mean}\left(LMP_i\right)\right|}{\text{mean}\left(LMP_i\right)} \geq p, \tag{14}$$

where mean(LMP_i) is the mean value of LMP at Bus i, which is obtained from historical data. In this work, we set p as 50%.

Step 3: LMP Calculation and EPRA. Based on equation (13), if the sample meets the adaptability criterion, the LMP can be calculated by solving a set of linear equations; otherwise, we should run DC-OPF to obtain the LMP. Then, with all the LMP data in hand, the EPRA can be completed by performing statistical analysis on the LMP data. Note that the EPRA results are obtained by the Monte Carlo (MC) simulation, which generates massive random samples that are carried out on the OPF model. The proposed method provides an effective tool to replace the time-consuming OPF calculation process to reduce the computationally demanding of the MC simulation. Hence, the convergence of the proposed method is the same as the traditional MC.

3. GP Surrogate Model for DC-OPF

This section first briefly introduces the GP. Then, the GP surrogate model for DC-OPF is proposed to identify the marginal generators and the congested transmission lines. The basic idea of the proposed GP surrogate model is to use the GP to replace the time-consuming optimization process of DC-OPF, as illustrated in Figure 2. The complicated DC-OPF features can be represented by the mapping relationship f: $D_i \longrightarrow P_i$, PF_l, which is the learning target of the GP.

3.1. A Brief Introduction to the GP Regression Method. The GP is generally used to solve hard regression and classification problems. It is attractive because of its flexible nonparametric nature and computational simplicity. In nonparametric statistics, the regularity of a relationship can be postulated without requiring the dataset to be focused on an easily describable class. This efficient property allows the GP to predict the functional behavior inside and outside of the input domain with a small sample size [35].

The GP is introduced for regression in this paper. We denote the regression function by $f(\cdot)$, which is the output of the GP surrogate model. Its corresponding input vector of p dimensions is denoted as $\mathbf{x}$. For a GP regression problem, a finite collection of training sample inputs $\boldsymbol{x}$ is denoted as $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$. Accordingly, the corresponding output $f(\mathbf{x})$ can be denoted as $[f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)]$. According to [36], the model output $f(\mathbf{x})$ is expected to follow a joint multivariate normal probability distribution, as follows:

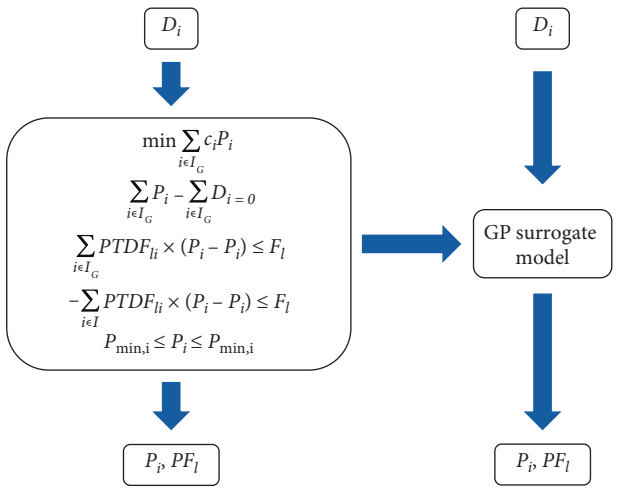

Figure 2: Relationship between DC-OPF and the GP.

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \begin{bmatrix} C(\mathbf{x}_1,\mathbf{x}_1) & \cdots & C(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ C(\mathbf{x}_n,\mathbf{x}_1) & \cdots & C(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix} \right), \quad (15)$$

where $m(\cdot)$ represents the mean function and $C(\cdot,\cdot)$ is a kernel function representing the covariance function. Then, (15) can be rewritten as

$$\mathbf{f}(\mathbf{X})|\mathbf{X} \sim \mathcal{N}(m(\mathbf{X}), C(\mathbf{X},\mathbf{X})), \quad (16)$$

where X is an $n\times p$ matrix denoted by $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^{\mathrm{T}}$. Then, considering that there is independent, identically distributed noise in the model output $\mathbf{f}(\mathbf{X})$, the realizations $\mathbf{Y}$ can be formulated as

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}\left(m(\mathbf{X}), C(\mathbf{X},\mathbf{X}) + \sigma^2 \mathbf{I}_n\right), \quad (17)$$

where σ^2 and $\mathbf{I}_n$ are the variances of the noise and an n-dimensional identity matrix, respectively. Note that the noise is always accounted for in practical implementation in the GP output.

To infer the GP regression output with noise from the abovementioned sample set ($\mathbf{X}$, $\mathbf{Y}$), a Bayesian inference framework is introduced. It is well known that a Bayesian posterior distribution of the model output can be inferred from a Bayesian prior distribution of $y(\mathbf{x})|\mathbf{x}$ and the likelihoods obtained from the realizations. For a test sample input set $\mathbf{X}_t$, the Bayesian prior distribution of $\mathbf{Y}_t$ can be expressed as

$$\mathbf{Y}_t|\mathbf{X}_t \sim \mathcal{N}\left(m(\mathbf{X}_t), C(\mathbf{X}_t,\mathbf{X}_t) + \sigma^2 \mathbf{I}_{n_t}\right). \quad (18)$$

Combined with the training sample set ($\mathbf{X}$, $\mathbf{Y}$), the joint distribution of $\mathbf{Y}$ and $\mathbf{Y}_t|\mathbf{X}$ can be formulated as follows:

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}_t|\mathbf{X} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{X}) \\ m(\mathbf{X}_t) \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \right), \quad (19)$$

where $\mathbf{C}_{11} = \mathbf{C}(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}_n$, $\mathbf{C}_{12} = \mathbf{C}(\mathbf{X}, \mathbf{X}_t)$, $\mathbf{C}_{21} = \mathbf{C}(\mathbf{X}_t, \mathbf{X})$, and $\mathbf{C}_{22} = \mathbf{C}(\mathbf{X}_t, \mathbf{X}_t) + \sigma^2\mathbf{I}_{nt}$. Then, based on the rules of the conditional Gaussian distribution, the Bayesian posterior distribution of $\mathbf{Y}_t$ can be inferred from the training sample set ($\mathbf{Y}$, $\mathbf{X}$) and test sample input set $\mathbf{X}_t$. It follows a Gaussian distribution $N(\mu(\mathbf{X}_t), \Sigma(\mathbf{X}_t))$. The expected value of $\mathbf{Y}_t$ can be expressed as follows:

$$\mu(\mathbf{X}_t) = m(\mathbf{X}_t) + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}(\mathbf{Y} - m(\mathbf{X})). \quad (20)$$

Thus far, the general form of GP regression has been derived. Equation (20) can be used as a surrogate model for a complicated DC-OPF model with a low computational cost. Further details about the GP can be found in [35, 36].

3.2. Proposed GP Surrogate Model for DC-OPF. The basic DC-OPF model introduced in Section 2 can be further expressed as the following linear programming (LP) problem:

$$\begin{aligned} &\min c(\mathbf{y}) \\ &\text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \ge \mathbf{b}, \mathbf{y} \in \Omega, \end{aligned} \quad (21)$$

where $\mathbf{y}$ is the output set, including the generation output and transmission line power flow, which are the key variables for determining the marginal generators and congested transmission lines, and c(·) is the associated cost function. $\mathbf{A}$ and $\mathbf{B}$ are the corresponding matrices concerning vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. With a random $\mathbf{x}$, the DC-OPF in (21) can be cast as a POPF problem.

In the proposed approach, a GP surrogate model is developed for the DC-OPF problem with a lower computational cost to improve the effectiveness of POPF. To this end, the steps for the proposed GP surrogate method are illustrated below.

3.2.1. Training Sample Generation. To construct the GP surrogate model described in (20) for the DC-OPF problem, the training sample set $\mathbf{D}$ = ($\mathbf{X}$, $\mathbf{Y}$) can be obtained from historical operational ISO data or by running an MC simulation where the uncertainty vector ω is sampled and the resulting DC-OPF output is calculated for a large number of samples. In particular, the Latin hypercube sampling method is implemented due to the small sample requirement of the GP. Here, each row of $\mathbf{X}$ is an I-dimensional uncertain input vector, including the load demands D_i of all buses. The output matrix $\mathbf{Y}$ contains columns of $I_G + I_L$ output variables as the generator output $\mathbf{P}$ and transmission line power flow $\mathbf{PF}$ corresponding to each input vector $\mathbf{x}$.

3.2.2. GP Surrogate Model Construction. With the training dataset $\mathbf{D} = [\mathbf{X}, \mathbf{Y}]$, we choose the squared exponential (SE) covariance kernel function for our regression problem, i.e.,

$$C_{SE}(\mathbf{x}_k, \mathbf{x}_*) = \tau^2 \exp\left(-\frac{(\mathbf{x}_k - \mathbf{x}_*)^T(\mathbf{x}_k - \mathbf{x}_*)}{2l^2}\right). \quad (22)$$

The hyperparameters $\boldsymbol{\xi} = (\tau, l)$ can be estimated by the Gaussian maximum likelihood estimator (MLE) method, which is optimal under the Gaussian assumption and is easy to implement [31]. Equation (17) with hyperparameters can be rewritten as

$$\mathbf{Y}|\mathbf{X},\boldsymbol{\xi} \sim N\left(m(\mathbf{X}), C_{SE}(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right). \tag{23}$$

Based on the MLE, we obtain

$$(\widehat{\xi},\widehat{\sigma}) = \arg\max_{\xi,\sigma} \log P(\mathbf{Y}|\mathbf{X},\boldsymbol{\xi},\sigma). \tag{24}$$

The marginal log-likelihood can be expressed as

$$\begin{aligned}\log P(\mathbf{Y}|\mathbf{X},\boldsymbol{\xi},\sigma) &= -\frac{1}{2}(\mathbf{Y}-m(\mathbf{X}))^T\left[C(\mathbf{X},\mathbf{X}|\boldsymbol{\xi}) + \sigma^2\mathbf{I}_n\right]^{-1}(\mathbf{Y}-m(\mathbf{X})) \\ &\quad -\frac{n}{2}\log 2\,\pi - \frac{1}{2}\log\left|C(\mathbf{X},\mathbf{X}|\boldsymbol{\xi}) + \sigma^2\mathbf{I}_n\right|.\end{aligned} \tag{25}$$

After utilizing a gradient-based optimizer, the hyperparameters are obtained while the GP surrogate model for DC-OPF is fully constructed.

3.2.3. The Key Information Identification for EPRA. For the output of the proposed GP surrogate model (e.g., **P** and **PF**), the learning error is not avoided. To identify the key information for EPRA (e.g., the marginal generator and congested transmission line) and address the learning error effect, a relaxing factor ε is introduced. Then, the marginal generators and congested transmission lines can be identified according to the following equations:

Marginal generator i:

$$P_{\min,i} + \varepsilon_{g,i} \le P_i \le P_{\max,i} - \varepsilon_{g,i}. \tag{26}$$

Congested transmission line l:

$$\left(PF_l \ge F_l - \varepsilon_{PF,l} \text{or} PF_l \le -F_l + \varepsilon_{PF,l}\right). \tag{27}$$

After obtaining the marginal generators and congested transmission lines of each sample, based on the framework proposed in Section 2, the EPRA can be achieved in short order without accuracy loss.

4. Numerical Results

In this section, the proposed method is tested on the IEEE 30-bus system to illustrate its effectiveness, while the IEEE 118-bus system is used to demonstrate its scalability to the larger system. All simulations are performed on a PC equipped with an AMD Ryzen 5 3600X 6-Core Processor CPU @ 3.80 GHz with 16 GB RAM. The algorithm is implemented in MATLAB.

The following methods are compared:

(i) M0: Monte Carlo simulation (benchmark)

(ii) M1: a neural network method based on SAE [37]

(iii) M2: a neural network method based on SDAE [38]

(iv) M3: a stacked extreme learning machine [24]

(v) M4: the Gaussian process [36]

(vi) M5: the proposed method

The hyperparameter settings of each data-driven method are shown in Table 1, which are obtained according to the artificial experience and reference [39]. The learning error of M1~M5, which is defined as the average difference between the results obtained with M1~M5 and M0, is evaluated as follows:

$$\begin{aligned}\Delta_1 &= \frac{\sum_{i=1}^{K}\sum_{j=1}^{p}\left|\widehat{y}_{i,j} - y_{i,j}\right|}{K\cdot p\cdot\sum_{i=1}^{K}\sum_{j=1}^{p}\left|y_{i,j}\right|}\times 100\%, \\ \Delta_2 &= \frac{1}{K\cdot p}\sum_{i=1}^{K}\sum_{j=1}^{p}\left|\frac{\widehat{y}_{i,j} - y_{i,j}}{y_{i,j}}\right|\times 100\%,\end{aligned} \tag{28}$$

where $\widehat{y}_{i,j}$ and p are the output of the data-driven method and its dimensions; $y_{i,j}$ is the output of Monte Carlo simulation as benchmark; K represents the number of testing samples; and Δ_2 is the MAPE index.

4.1. Evaluation of the Proposed Approach on the IEEE 30-Bus System

4.1.1. Learning Performances of DC-OPF and LMP. We first show that the EPRA problem of LMP assessment is more complicated than the OPF problem, which cannot be learned directly by data-driven methods. The learning performances of the LMP and DC-OPF outputs learned by M1–M4 are compared on IEEE 30 in Table 2. For M1–M3, the number of training samples is set as 10000. For M4 and M5, the number of training is set as 200. The number of testing samples is set as 10000 for all methods.

The results show that directly learning the LMP is intractable for data-driven methods because of its discontinuous property. Additionally, the DC-OPF problem is more comfortable to learn, making the proposed method based on the learning output of DC-OPF reasonable.

4.1.2. Effectiveness of the Proposed Method. To demonstrate the benefits achieved by the proposed approach, we compare the LMP errors of M1–M5 in the IEEE 30-bus system, as shown in Table 3.

Several conclusions can be drawn, as follows:

(1) Among all the methods for EPRA, the proposed method (M5) achieves the best accuracy, which is exactly the same as that of the benchmark method. In the proposed framework, the learning error of the GP is filtered out by the identification process and model-based adaptability criterion.

(2) Compared with M0, the testing time decreases by 59.34%. This shows that the computational efficiency of LMP is significantly improved by the proposed method without accuracy loss.

(3) Comparing M4 with M1–M3, the results show that the GP can achieve a similar accuracy with a small sample size. However, the learning errors of the data-driven methods are unavoidable, even with a large number of training samples.

Table 1: Hyperparameter settings of the data-driven methods.

Case	Method	Hyperparameter settings
IEEE 30	M1	3 layers, 100 nodes per layer, and 200 epochs; learning rate = 0.0001, branch size = 500
	M2	3 layers, 100 nodes per layer, and 200 epochs; learning rate = 0.0001, branch size = 500
	M3	500 nodes, 50 reduced hidden nodes, and 2 epochs
	M4	$m(x) = 0$, $C(\cdot, \cdot) = C_{SE}(\cdot, \cdot)$, 100 epochs
	M5	$m(x) = 0$, $C(\cdot, \cdot) = C_{SE}(\cdot, \cdot)$, 100 epochs
IEEE 118	M1	3 layers, 300 nodes per layer, and 300 fine-tuning epochs; learning rate = 0.0001; branch size = 500
	M2	3 layers, 300 nodes per layer, and 300 fine-tuning epochs; learning rate = 0.0001; branch size = 100
	M3	1000 nodes, 100 reduced hidden nodes, and 4 epochs
	M4	$m(x) = 0$, $C(\cdot, \cdot) = C_{SE}(\cdot, \cdot)$, 100 epochs
	M5	$m(x) = 0$, $C(\cdot, \cdot) = C_{SE}(\cdot, \cdot)$, 100 epochs

Table 2: Average error comparison between LMP and DC-OPF on the IEEE 30-bus system

Method	Δ_2 of LMP (%)	Δ_1 of DC-OPF outputs	
		P (%)	**PF** (%)
M1	5.98	2.71	5.10
M2	5.93	2.73	5.57
M3	5.95	1.76	2.72
M4	7.11	2.01	3.32

Table 3: LMP average error comparison on the IEEE 30-bus system

Method	Number of training samples	Training time (s)	Testing time (s)	Δ_1 (%)	Δ_2 (%)
M0	—	—	150.01	—	—
M1	10000	11.72	0.021	6.30	5.98
M2	10000	21.56	0.025	6.41	5.93
M3	10000	1.01	0.34	6.17	5.95
M4	200	7.48	1.38	7.53	7.11
M5	**200**	**10.41**	**60.99**	**0**	**0**

Bold values are used for highlighting the results of the proposed method in this paper.

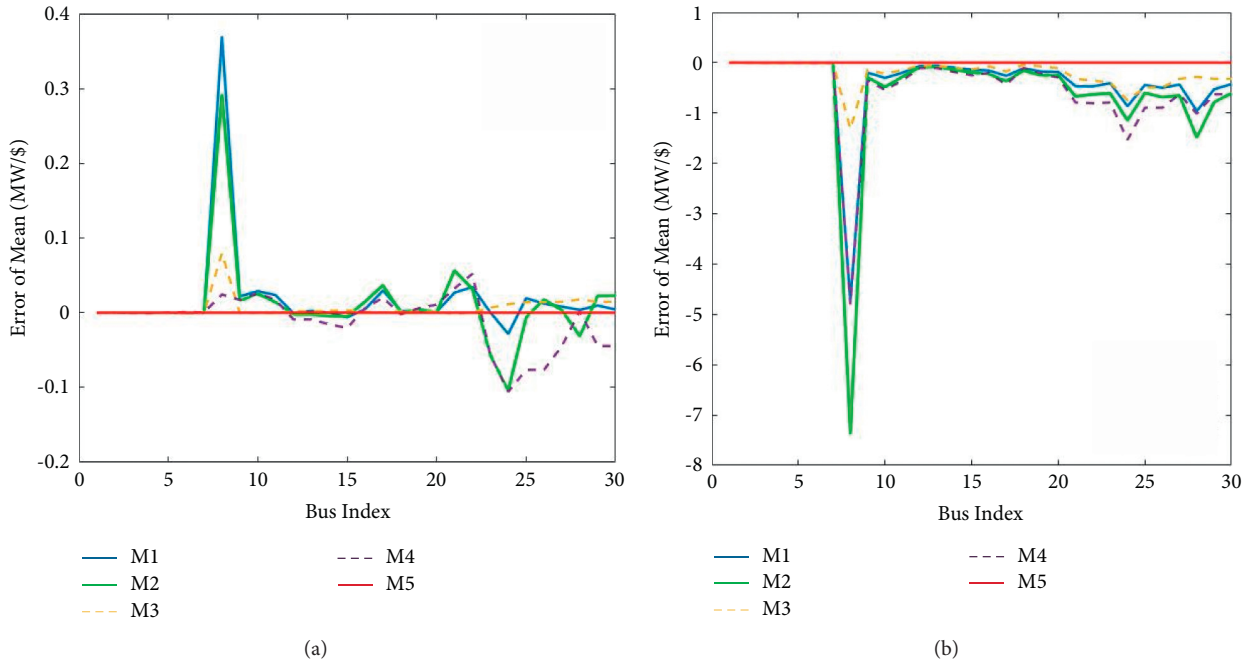

Figure 3: EPRA error comparison on the IEEE 30-bus system. (a) Error of the mean. (b) Error of the standard deviation.

Table 4: LMP average error comparison on the IEEE 118-bus system.

Method	Number of training samples	Training time (s)	Testing time (s)	Δ_1 (%)	Δ_2 (%)
M0	—	—	227.04	—	—
M1	10000	94.52	0.07	4.13	3.60
M2	10000	123.51	0.10	3.76	3.43
M3	10000	1.17	0.38	8.39	6.55
M4	200	17.52	6.88	4.30	4.87
M5	**200**	**49.89**	**123.27**	**0**	**0**

Bold values are used for highlighting the results of the proposed method in this paper.

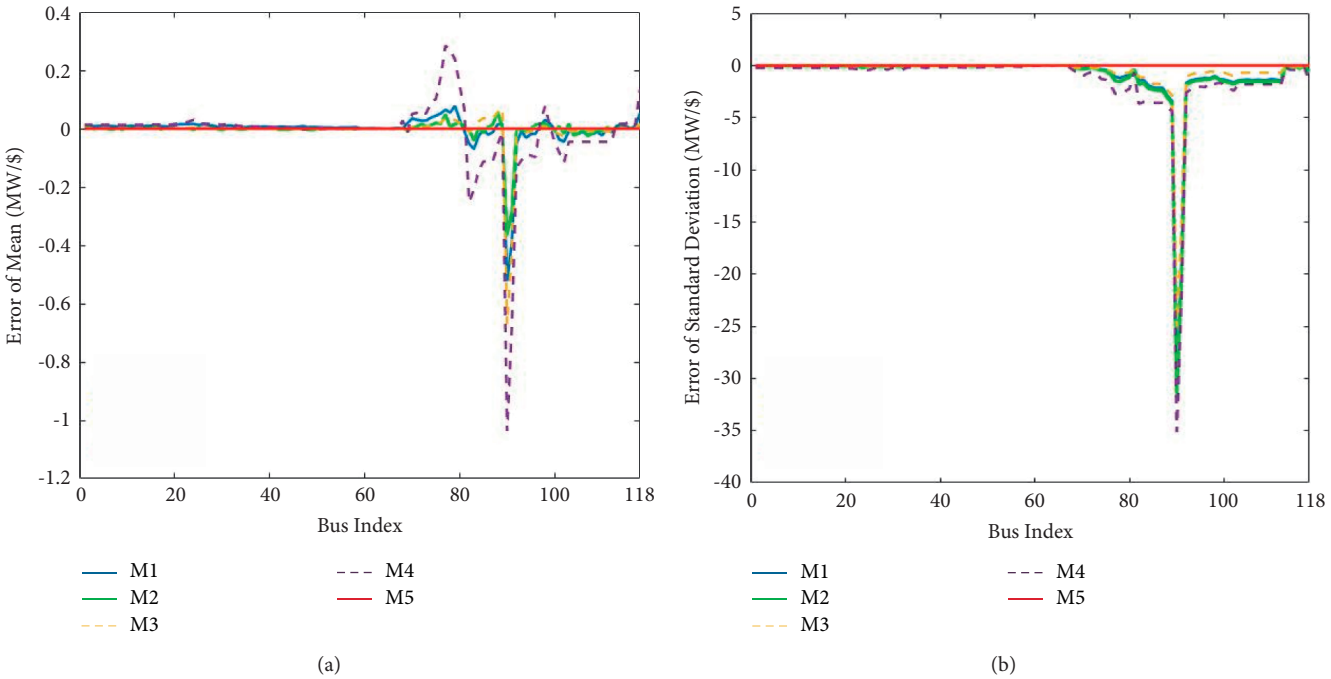

Figure 4: EPRA result comparison on the IEEE 118-bus system. (a) Error of the mean. (b) Error of the standard deviation.

(4) For M1–M5, because of the superior quality of the GP (e.g., its small sample requirement), the training sample size of the proposed method is much smaller than those of M1–M3, which aligns well with the current industry practice.

The errors of the EPRA results are compared, as shown in Figure 3. For the proposed method, the EPRA results for the mean and standard deviation are very accurate. However, for M1~M4, the estimation results for the mean are relatively accurate, while the standard deviations are far from the real value obtained by M0, so they cannot achieve an accurate EPRA. In particular, as seen in Figure 3, the error is abnormally large in some specific buses (e.g., Bus 8 and Bus 24). This is because the LMPs at these buses fluctuate much more than other buses, making it challenging to be directly learned by the data-driven methods, resulting in false information being passed to ISOs and market participants, leading to severe market risks.

4.2. Results on the IEEE 118-Bus System. The IEEE 118-bus system is used to demonstrate the scalability of the proposed method. The learning error and EPRA results are given in Table 4 and Figure 4, respectively. The test sample number is set to be the same as for the IEEE 30-bus system.

The results show that the proposed method can guarantee EPRA accuracy even in a large case while improving the computational efficiency. It also shows that the statistical moments (i.e., the mean and standard deviation) diverge far from the real value at Buses 72~110, which demonstrates that EPRA cannot be achieved by data-driven direct learning.

5. Conclusions and Future Work

This paper proposes a data-driven framework for EPRA. Specifically, a GP surrogate model is developed to identify the marginal generators and congested transmission lines of DC-OPF. This paves the way for improving the efficiency of EPRA. Based on the KKT condition, an adaptability criterion is proposed to identify samples with learning errors. The simulation results show that the proposed method increases EPRA accuracy. Comparisons with recent data-driven methods show that the proposed approach can greatly improve the computational efficiency of EPRA without compromising its accuracy. It is also shown that direct

learning for LMP may not be tractable due to the problem complexity. Future work will consider more uncertain boundary conditions for the power system.

As shown in this paper, although the inherent learning error in data-driven methods is unavoidable, helpful reference information can be analyzed to increase the computational speed of EPRA. Hence, the idea of improving the efficiency of spot market risk analysis through data-driven techniques is worthy of further study. This paper focuses on the EPRA in the spot market, which is commonly used in the Guangdong Electric Power Trading Center of China to manage LMP risk. Our study shows that the GP is capable of learning the complex relationship between a set of key information (e.g., the marginal generator and congested transmission line) and system operating conditions with minimal training samples and a fast training speed. Therefore, for more complicated problems considering bid price, generating capacity, and unit commitment, the GP is also a promising tool for extracting useful information from historical data. However, effectively utilizing GP techniques regarding the specific properties of these problems is worthy of further exploration. In addition, the physical models of spot market operation are known by the ISO. Therefore, improving the learning performance in combination with power domain expertise should be investigated further. To further speed up the calculation, it is necessary to improve the learning accuracy considering the characteristics of the physical model and to make the proposed method applicable to more scenarios according to the selection of samples in future work.

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project was supported by the Science and Technology Project of Guangdong Electric Power Trading Center Co., Ltd. (GDKJXM20185365).

References

[1] J. Shi, Z. Ding, W. J. Lee, Y. Yang, Y. Liu, and M. Zhang, "Hybrid forecasting model for very-short term wind power forecasting based on grey relational analysis and wind speed distribution features," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 521–526, 2014.

[2] M. Zhu, C. Xu, S. Dong, K. Tang, and C. Gu, "An integrated multi-energy flow calculation method for electricity-gas-thermal integrated energy systems," *Protection and Control of Modern Power Systems*, vol. 6, no. 1, pp. 65–76, 2021.

[3] M. Erdiwansyah, H. Husin, M. Zaki, and Muhibbuddin, "A critical review of the integration of renewable energy sources with various technologies," *Protection and Control of Modern Power Systems*, vol. 6, no. 1, pp. 37–54, 2021.

[4] J. Huang, Y. Xue, Z. Y. Dong, and K. P. Wong, "An efficient probabilistic assessment method for electricity market risk management," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1485–1493, 2012.

[5] R. Yao, S. Huang, K. Sun et al., "Risk assessment of multi-timescale cascading outages based on markovian tree search," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2887–2900, 2017.

[6] R. Wang, Q. Sun, P. Tu, J. Xiao, Y. Gui, and P. Wang, "Reduced-order aggregate model for large-scale converters with inhomogeneous initial conditions in DC microgrids," *IEEE Transactions on Energy Conversion*, vol. 36, no. 3, pp. 2473–2484, 2021.

[7] R. Wang, Q. Sun, W. Hu, Y. Li, D. Ma, and P. Wang, "SoC-Based droop coefficients stability region analysis of the battery for stand-alone supply systems with constant power loads," *IEEE Transactions on Power Electronics*, vol. 36, no. 7, pp. 7866–7879, 2021.

[8] P. Malo, "Modeling electricity spot and futures price dependence: a multifrequency approach," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 22, pp. 4763–4779, 2009.

[9] A. J. Conejo and M. Carrion, "Risk constrained electricity procurement for a large consumer," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 153, no. 4, pp. 407–413, 2006.

[10] R. Dahlgren, C. C. Chen-Ching Liu, and J. Lawarree, "Risk assessment in energy trading," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 503–511, 2003.

[11] M. Liu and F. F. Wu, "Managing price risk in a multimarket environment," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1512–1519, 2006.

[12] Y. Ding and P. Wang, "Reliability and price risk assessment of a restructured power system with hybrid market structure," *IEEE Transactions on Power Systems*, vol. 21, no. 1, pp. 108–116, Feb. 2006.

[13] Y. Wang, Z. Yang, J. Yu, and X. Fang, "Revisit the electricity price formulation: a formal definition, proofs, and examples," *Energy*, vol. 200, 2020.

[14] G. Verbic and C. A. Canizares, "Probabilistic optimal power flow in electricity markets based on a two-point estimate method," *IEEE Transactions Power System*, vol. 21, no. 4, pp. 1883–1893, 2006.

[15] G. Sun, S. Chen, Z. Wei, and S. Chen, "Multi-period integrated natural gas and electric power system probabilistic optimal power flow incorporating power-to-gas units," *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 3, pp. 412–423, 2017.

[16] C.-L. Su, "Probabilistic load-flow computation using point estimate method," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 1843–1851, 2005.

[17] A. Schellenberg, W. Rosehart, and J. Aguado, "Introduction to cumulant-based probabilistic optimal power flow (P-OPF)," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 1184–1186, 2005.

[18] J. Malinowski, "A monte carlo method for estimating reliability parameters of a complex repairable technical system with inter-component dependencies," *IEEE Transactions on Reliability*, vol. 62, no. 1, pp. 256–266, 2013.

[19] S. Wang, Y. Ding, C. Ye, C. Wan, and Y. Mo, "Reliability evaluation of integrated electricity–gas system utilizing network equivalent and integrated optimal power flow

techniques," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 6, pp. 1523–1535, 2019.

[20] Y. Li, W. Gao, W. Yan et al., "Data-driven optimal control strategy for virtual synchronous generator via deep reinforcement learning approach," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 919–929, 2021.

[21] Y. Yang, Z. Yang, J. Yu, K. Xie, and L. Jin, "Fast economic dispatch in smart grids using deep learning: an active constraint screening approach," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11030–11040, 2020.

[22] Y. Yang, J. Yu, Z. Yang, M. Xiang, and R. Liu, "Fast calculation of probabilistic optimal power flow: a deep learning approach," in *Proceedings of the IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, Atlanta, GA, USA, August 2019.

[23] X. Pan, T. Zhao, M. Chen, and S. Zhang, "DeepOPF: a deep neural network approach for security-constrained DC optimal power flow," *IEEE Transaction Power System*, vol. 36, no. 3, pp. 1725–1735, 2021.

[24] X. Lei, Z. Yang, J. Yu, J. Zhao, Q. Gao, and H. Yu, "Data-driven Optimal power flow: a physics-informed machine learning approach," *IEEE Transaction Power System*, vol. 36, no. 1, pp. 346–354, 2021.

[25] Y. Yang, Z. Yang, J. Yu, B. Zhang, Y. Zhang, and H. Yu, "Fast Calculation of probabilistic power flow: a model-based deep learning approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2235–2244, 2020.

[26] G. B. Huang and L. Chen, "Siew universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879–892, 2006.

[27] J. Feng, X. Jia, H. Cai, F. Zhu, X. Li, and J. Lee, "Cross trajectory gaussian process regression model for battery health prediction," *Journal of Modern Power System Clean Energy*, vol. 9, no. 5, pp. 1217–1226, 2021.

[28] J. Han, X. P. Zhang, and F. Wang, "Gaussian process regression stochastic volatility model for financial time series," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1015–1028, 2016.

[29] H. Liu, Y. S. Ong, and J. Cai, "A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design," *Structural and Multidisciplinary Optimization*, vol. 57, no. 1, pp. 393–416, 2018.

[30] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: a review," *Foundations and Trends in Machine Learning*, vol. 4, no. 3, pp. 195–266, 2012.

[31] H. Liu and J. Cai, "Ong. remarks on multi-output gaussian process regression," *Knowledge -Based System*, vol. 144, pp. 102–121, 2018.

[32] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *Journal Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[33] B. Shahriari, K. Swersky, Z. Wang, N. de Freitas, and R. P. de Freitas, "Taking the human out of the loop: a review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[34] Z. Yang, A. Bose, H. Zhong et al., "LMP revisited: a linear model for the loss-embedded LMP," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 4080–4090, 2017.

[35] H. Liu, Y. S. Ong, X. Shen, and J. Cai, "When gaussian process meets big data: a review of scalable GPs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4405–4423, 2020.

[36] Y. Xu, Z. Hu, L. Mili, M. Korkali, and X. Chen, "Probabilistic power flow based on a gaussian process emulator," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 3278–3281, 2020.

[37] C. Lu, Z. Y. Wang, W. L. Qin, and J. Ma, "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Processing*, vol. 130, pp. 377–388, 2017.

[38] C. Xing, L. Ma, and X. Yang, "Stacked denoise autoencoder based feature extraction and classification for hyperspectral images," *Journal of Sensors*, vol. 2016, Article ID 3632943, 10 pages, 2016.

[39] Q. Gao, Z. Yang, J. Yu et al., "Model-driven architecture of extreme learning machine to extract power flow features," *IEEE Transaction Neural Network Learning System*, vol. 32, no. 10, pp. 4680–4690, 2021.