Chance Constrained Extreme Learning Machine for Nonparametric Prediction Intervals of Wind Power Generation

Can Wan, *Member, IEEE*, Changfei Zhao, and Yonghua Song, *Fellow, IEEE*

***Abstract*—Confronted with considerable intermittence and variability of wind power, prediction intervals (PIs) serve as a crucial tool to assist power system decision-making under uncertainties. Conventional PIs rely on predetermining the lower and upper quantile proportions and therefore suffer from conservative interval width. This paper innovatively develops a chance constrained extreme learning machine (CCELM) model to generate quality nonparametric proportion-free PIs of wind power generation, which minimizes the expected interval width subject to the PI coverage probability constraint. Due to the independency on the preset PI bounds proportions, the proposed CCELM model merits high adaptivity and taps the latent potentialities for PI shortening. The convexity of extreme learning machine renders the sample average approximation counterpart of stochastic CCELM model equivalent to a parameter searching task in parametric optimization problem with polyhedral feasible region. A novel difference of convex functions optimization based bisection search (DCBS) algorithm is proposed to efficiently construct the CCELM model, which successfully realizes machine learning by means of solving linear programming problems sequentially. Comprehensive numerical experiments based on actual wind farm data demonstrate the significant effectiveness and efficiency of the developed CCELM model and DCBS algorithm.**

***Index Terms*—Prediction interval, forecasting, wind power, chance constraint, extreme learning machine, DC optimization.**

NOMENCLATURE

A. Acronyms

ACD	Average coverage deviation.
AW	Average width.
BB	Branch-and-bound.
CCELM	Chance constrained extreme learning machine.
DC	Difference of convex functions.
DCA	DC optimization algorithm.
DCBS	DC optimization based bisection search.
ECP	Empirical coverage probability.
ELM	Extreme learning machine.
LP	Linear programming.
MILP	Mixed integer linear programming.
NCP	Nominal coverage probability.
PDF	Probability density function.
PI	Prediction interval.
PO	Parametric optimization.
PZO	Parametric zero-one loss optimization.
SAA	Sample average approximation.

B. Parameters and Criteria

$\boldsymbol{a}_j$	Weight vector of ELM linking the input layer to the jth hidden-layer neuron.
b_j	Bias of ELM for the jth hidden-layer neuron.
v	Budget for the overall interval width.
$v^\star$	Optimal value of the SAA problem.
$\hat{v}$, $\underline{v}$	Goal parameter values for the parameter searching tasks in the original and relaxed PZO problems.
$\underline{v}_\ell$, $\underline{v}_u$	Lower and upper bounds of the searching range for the goal parameter value.
$\overline{v}$	Goal parameter value returned by the DCBS algorithm.
I, H, O	Neuron numbers for the input, hidden, and output layers of ELM.
S_β	Interval score for central PIs with $100(1-\beta)\%$ confidence level.
$\underline{\alpha}$, $\overline{\alpha}$	Lower and upper quantile proportions of PIs.
$1-\beta$	Nominal coverage probability of PIs.
ϵ_1, ϵ_2	Tolerances for bisection search and DCA.

C. Variables and Functions

$\boldsymbol{f}$, f	Output functions for the whole output layer and one of the output-layer neurons of ELM.
$\boldsymbol{h}$	Hidden-layer output function of ELM.
ℓ_t, u_t	Lower and upper bounds of PI indexed by t.
$\boldsymbol{x}$, $\boldsymbol{x}_t$	Random variable for input feature vector, and its realization indexed by t.

Manuscript received August 18, 2019; revised December 29, 2019 and February 28, 2020; accepted March 29, 2020. Date of publication April 14, 2020; date of current version August 24, 2020. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB0905000, in part by the National Natural Science Foundation of China under Grants 51877189 and 51761135015, and in part by the Young Elite Scientists Sponsorship Program by the China Association of Science and Technology under Grant 2018QNRC001. The work of C. Wan was supported by the Hundred Talents Program of Zhejiang University. Paper no. TPWRS-01225-2019. *(Corresponding author: Can Wan.)*

Can Wan and Changfei Zhao are with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: canwan@zju.edu.cn; zhaochangfei@zju.edu.cn).

Yonghua Song is with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Taipa, Macau SAR, China, and also with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: yhsong@um.edu.mo).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2020.2986282

y, y_t	Random variable for wind power generation, and its realization indexed by t.
L_{H}, L_{T}, L_{DH}	Hinge loss, truncated hinge loss, and difference of hinge variants.
L^{+}_{vex}, L^{-}_{vex}	Convex minuend and subtrahend of DC.
$\boldsymbol{\gamma}$, γ_t	Vector and its element indexed by t indicating the target hits or misses of PIs.
$\boldsymbol{\delta}$	Subgradient of nonsmooth convex function.
$\boldsymbol{\theta}$	Vector concatenating decision variables.
μ	Joint probability distribution of input features and targets.
ρ, $\underline{\rho}$, $\overline{\rho}$	Optimal value functions of the original PZO, relaxed PZO, and parametric DC optimization problems.
φ	Hidden-layer activation function of ELM.
$\boldsymbol{\omega}_{\ell}$, $\boldsymbol{\omega}_{u}$	Output weight vectors of ELM corresponding to the lower and upper bounds of PIs.
$\boldsymbol{\tau}$, τ_t	Auxiliary vector and its element indexed by t for nonsmooth optimization linearization.

D. Sets

$\hat{\mathcal{I}}_t^{(\beta)}$	PI with $100(1-\beta)\%$ confidence level indexed by t.
$\mathcal{T}$, $\mathcal{V}$	Index set of training and test datasets.
Ω_{SAA}	Feasible region of the SAA problem.
Ω_{PO}, $\overline{\Omega}_{\mathrm{PO}}$	Feasible regions of the original and relaxed PZO problems.

E. Operators

$\lvert\cdot\rvert$	Cardinality of set.
$\lVert\cdot\rVert_p$	p-norm of vector.
$\langle\cdot,\cdot\rangle$	Inner product of two vectors.
$\mathbb{E}[\cdot]$	Expectation operator.
$\mathbb{I}(\cdot)$	Indicator function.
$\mathbb{P}(\cdot)$	Probability operator.

I. Introduction

WIND energy is considered as the leading renewable resource for electricity generation with technology maturity and price competitiveness. Under continuously declining cost of facilities and maintenance, there was 51.3 GW new global capacity installation of wind power in 2018, which nearly contributes to 30% of the total renewable growth [1]. However, wind power uncertainties originated from the chaotic meteorological systems have brought severe risks and challenges to wind energy integrations. Therefore, accurate wind power forecasting is of vital importance to enhance the security and economy of operation and control in power systems.

Conventional deterministic point forecasting methods concentrate on a single predictive value of wind power, which neglects valuable probabilistic information and suffers from unavoidable forecasting deviations. By contrast, the emerging probabilistic forecasting methodologies produce a series of prediction values along with their probabilistic explanations to quantify the uncertainties of wind power. Meanwhile, modern power system operations under uncertainties rely on probabilistic forecasting, such as maximum uncertainty boundary analysis of distribution systems [2], stochastic energy and reserve scheduling [3], stochastic receding horizon control of active distribution networks [4], power system balancing risk evaluation [5], etc. In this context, advanced probabilistic wind power forecasting approaches gain more attentions and are extensively studied in the literature.

Many pioneer works aim at quantifying the uncertainties of deterministic wind power forecasting. These methodologies firstly obtain deterministic forecasting errors and then analyze the statistical distributions. Parametric analysis of wind power forecasting errors prescribes specific distributions, such as Gaussian [6], Beta [7], and Lévy α-stable [8] distributions, where the parameters can be identified by least squares fitting or maximum likelihood estimation [9]. Besides, ensemble learning techniques are combined with some parametric distributional assumption to conduct probabilistic wind power forecasting. In [10], [11], the bootstrap based neural networks (BNN) are trained to estimate the true regression value and the variance of wind power forecasting uncertainty. A bootstrap based extreme learning machine (BELM) model is proposed to effectively quantify the wind power forecasting uncertainties from model misspecification and data noise, which avoids complicated gradient calculation and merits superb computational efficiency [12]. Nonparametric analysis of wind power forecasting errors does not require these distributional assumptions. Conditional prediction intervals (PIs) of wind power are constructed by means of adaptively resampling empirical forecasting errors [13]. With deterministic wind power predictions and weather uncertainty information as the input, radial basis function neural network and self-organized map are combined to generate predictive quantiles of wind power generation [14]. Copula theory is utilized to model the forecasting error distributions of multiple wind farms conditional on deterministic forecasting values [3]. In order to evaluate the performance of deterministic wind power forecasting, spline quantile regression of forecasting errors is conducted with respect to a range of proportions [15].

Actually, probabilistic wind power forecasting does not necessarily require the exogenous information of deterministic forecasting. More studies focus on directly predicting the probabilistic properties of wind power independent of deterministic forecasting errors [16]. In [17], predictive distributions of wind power are constructed based on weather ensemble predictions and kernel density estimations. By means of optimizing the performance metrics of PIs, a hybrid intelligent algorithm (HIA) is developed to obtain the optimal nonparametric PIs of wind power [18]. Time adaptive Nadaraya-Watson density estimator with heterogeneous kernels is applied to obtain the predictive probability distributions of wind power generation [19]. Facilitated by extreme learning machine (ELM), a direct quantile regression (DQR) model is efficiently trained via linear programming (LP) to obtain a set of predictive quantiles of wind power [20]. The sparse Bayesian learning (SBL), kernel density estimation, and Beta distribution fitting approaches are optimally combined to forecast the probability distributions of wind power production [21]. Different from traditional Gaussian kernel density estimator (GKDE), a fuzzy and adaptive diffusion-based kernel density estimator (FADiE) is proposed in order to construct PIs according to the predictive density

functions of the training targets [22]. Infinite Markov switching autoregression is proposed to model the wind power time series, and the posterior distributions of wind power are generated via Bayesian inference and sampling procedures [23]. Bidirectional long short-term memory deep learning architecture and copula approach are combined to obtain the multivariate probabilistic wind power forecasting [24]. A time-varying copula updating mechanism is presented to generate joint scenarios from the marginal non-Gaussian densities of renewable generation [25]. Besides, probabilistic forecasting is also employed to estimate the forecasting uncertainty associated with electricity load and price [26]. Ellipsoidal predictive regions of renewable generations and electricity prices are constructed with reliable probability guarantee and minimum volume [27]. Deep learning models are adaptively assembled via improved bagging and boosting techniques to generate the probabilistic low-voltage load forecasting [28].

Probabilistic forecasting in the form of prediction intervals is becoming increasingly prevalent, since prediction intervals allow direct applications to robust or interval optimizations in power systems [27]. Given the nominal confidence level, shorter PIs subject to the satisfaction of reliability are preferred in power system decision-making issues. Traditional central PIs have been widely attended in previous studies [10]–[13], [16], [18]–[24], which restrict the lower and upper bounds of PIs to be probabilistically centered around the medians [29]. Recently, a machine learning based linear programming (MLLP) model is proposed and the dispensable central property of PIs is validated by means of sensitivity analysis [30]. An adaptive bilevel programming (ABP) model is formulated with quantile regression as the follower and quantile proportion tuning as the leader to obtain the shortest well-calibrated wind power PIs [31]. Actually, PIs derived by predictive probability density distributions (PDFs), interval score optimization, or quantile regression inevitably rely on prescribing bounds proportions, which motivates this work to develop a proportion-free interval forecasting model capable of simultaneously ensuring the reliability and sharpness of PIs.

In this paper, a chance constrained extreme learning machine (CCELM) model is innovatively developed, which minimizes the expected interval width subject to the PI coverage probability constraint. The CCELM model complies with the aim of probabilistic forecasting that maximizes sharpness with well reliability as the prerequisite [32]. The expectation and probability operators in the stochastic CCELM model are replaced by their sample average approximation (SAA) counterparts, independent of specific distributions of input feature and target variables. Then the SAA problem is equivalently reformulated as a parameter searching task in the parametric zero-one loss optimization (PZO) problem. Owing to the linear mathematical formulation of ELM, the feasible region of the PZO problem can be transformed into a convex polyhedron. To overcome the computational difficulty in minimizing zero-one loss over convex polyhedral region, a surrogate loss function is elaborated and takes the form of difference of convex functions (DC). A DC optimization based bisection search (DCBS) algorithm is accordingly proposed to fulfill nonparametric machine learning by means of solving a finite number of convex LP problems efficiently. Comprehensive numerical studies under an actual wind farm verify the superior performance of the developed CCELM model for generating quality PIs of wind power generation with consideration of multiple confidence levels, look-ahead steps, seasonal patterns, and proportion pairs. The significant effectiveness of the proposed DCBS procedure for training the CCELM model is validated via systematic comparisons with the state-of-the-art algorithms for solving chance constrained stochastic programming problems.

Major contributions of this work are summarized below:

1) A novel chance constrained extreme learning machine model is developed for nonparametric PIs of wind power generation, which is independent of proportion selection and merits high adaptivity.
2) The SAA counterpart of the stochastic CCELM model is equivalently transformed into a parameter searching task in parametric zero-one loss optimization problem subject to convex polyhedral constraints.
3) A difference of convex functions is elaborated to act as the surrogate for the zero-one loss with high accuracy and improves computational tractability of the zero-one loss minimization problem.
4) A DC optimization based bisection search algorithm is proposed to efficiently construct the CCELM model, which only requires solving a sequence of linear programming problems.

The remainder of this paper is organized as follows. Section II introduces the basics of nonparametric PIs and links the chance constraint with PI formulations. Section III formally presents the CCELM model and derives its sample average approximation counterpart. Section IV reformulates the SAA counterpart as a parameter searching task and proposes the DCBS algorithm to fulfill machine learning. Section V conducts comprehensive numerical experiments to verify the superior effectiveness of the CCELM model and DCBS algorithm. Finally, the conclusion of this paper is drawn in Section VI.

II. Prediction Intervals and Chance Constraint

A. Nonparametric Prediction Intervals

For normalized wind power within the range of [0,1], its prediction interval is defined as a closed set $\hat{\mathcal{I}}_t^{(\beta)}$ that encloses the future wind power y_t indexed by t with *nominal coverage probability* (NCP), expressed as

$$\hat{\mathcal{I}}_t^{(\beta)} \in \{[\ell_t, u_t] \subseteq [0,1] \mid \mathbb{P}\left(\ell_t \le y_t \le u_t\right) = 100(1-\beta)\%\} \tag{1}$$

where ℓ_t and u_t are the lower and upper bounds of PI; β is the risk level; $100(1-\beta)\%$ is the confidence level of PI $\hat{\mathcal{I}}_t^{(\beta)}$, namely, the nominal coverage probability of PI $\hat{\mathcal{I}}_t^{(\beta)}$. It is evident that the PI $\hat{\mathcal{I}}_t^{(\beta)}$ of wind power is generally not unique, though under the same NCP [30].

The lower and upper bounds of a PI can be determined via a pair of quantiles $q_t^{\underline{\alpha}}$ and $q_t^{\overline{\alpha}}$ with proportions $\underline{\alpha}$ and $\overline{\alpha}$ respectively,

given by

$$\hat{\mathcal{I}}_t^{(\beta)} := [\ell_t, u_t] = \left[q_t^{\underline{\alpha}}, q_t^{\overline{\alpha}}\right] \tag{2}$$

As shown by the PI definition (1), the NCP requires that the lower and upper proportions $\underline{\alpha}$ and $\overline{\alpha}$ should meet the following rule

$$\overline{\alpha} - \underline{\alpha} = 1 - \beta \tag{3}$$

which prescribes that the coverage probability of PI equals its nominal value $100(1-\beta)\%$.

Conventionally, the lower and upper proportions $\underline{\alpha}$ and $\overline{\alpha}$ are usually determined according to empirical rules. Central PIs with symmetric proportions with respect to the medians are one of the most prevalent options. Non-central PIs can be obtained via conducting sensitivity analysis on the *probability mass bias* (PMB) [30], defined as

$$\text{PMB} := (1 - \overline{\alpha}) - \underline{\alpha} \tag{4}$$

The PMB (4) combines the NCP requirement (3) to determine a unique proportion pair. However, since the predictive probability distributions of wind power generation are time-variant with significant skewness, both central and non-central PIs with invariant proportion pair might suffer from conservative interval width.

B. Assessment Criteria of Prediction Intervals

Generally, *reliability* and *sharpness* serve as the primary and secondary properties of prediction intervals respectively. Besides, reliability and sharpness can be merged into *overall skill*. Given a series of actual wind power for test $\{y_t\}_{t\in\mathcal{V}}$ and the corresponding PIs $\{[\ell_t, u_t]\}_{t\in\mathcal{V}}$, the aforementioned three properties of PIs can be indicated by the following assessment criteria.

1) Reliability: Reliability of PIs emphasizes the fidelity of their empirical coverage probability (ECP) to the NCP $100(1-\beta)\%$, which can be measured by the average coverage deviation (ACD), defined by

$$\text{ACD} := \text{ECP} - (1 - \beta) \tag{5}$$

The ECP is given by

$$\text{ECP} := \frac{1}{|\mathcal{V}|} \sum_{t\in\mathcal{V}} \mathbb{I}\left(\ell_t \le y_t \le u_t\right) \tag{6}$$

where $\mathcal{V}$ is the index set of test samples and $|\cdot|$ denotes the cardinality of set. The magnitude of ACD should reach zero as close as possible. In general, smaller absolute value of ACD indicates higher reliability of PIs.

2) Sharpness: Sharpness represents the concentration of PIs, which can be quantified via the average width (AW) of PIs, defined as

$$\text{AW} := \frac{1}{|\mathcal{V}|} \sum_{t\in\mathcal{V}} \left(u_t - \ell_t\right) \tag{7}$$

Since the interval width implies the uncertain degrees identified by forecasters, PIs with smaller AW are deemed more accurate and valuable on the premise of high reliability.

3) Overall Skill: Proper scoring rules are crucial to the overall quality assessment of PIs. The widely adopted negatively oriented interval score [32] is formulated as

$$S_\beta := \frac{1}{|\mathcal{V}|} \sum_{t\in\mathcal{V}} \Big[\left(u_t - \ell_t\right) + \frac{2}{\beta}\left(\ell_t - y_t\right)\mathbb{I}\left(y_t < \ell_t\right) + \frac{2}{\beta}\left(y_t - u_t\right)\mathbb{I}\left(y_t > u_t\right)\Big] \tag{8}$$

However, it should be highlighted that such score is derived based on central PIs and prejudiced against non-central and proportion-free PIs [30], [31].

C. Link Between Chance Constraint and Prediction Intervals

The goal of probabilistic forecasting is to maximize sharpness subject to reliability [32]. From the perspective of prediction intervals, sharpness and reliability can be indicated by the interval width and coverage probability respectively. In this regard, prediction intervals of smaller width and well reliability would be favored [33]. Therefore, the following chance constrained decision-making framework is established and expressed as

$$\min_{\ell_t, u_t} \quad u_t - \ell_t \tag{9a}$$

$$\text{s.t.} \quad \mathbb{P}\left(\ell_t \le y_t \le u_t\right) \ge 1 - \beta \tag{9b}$$

$$0 \le \ell_t \le u_t \le 1 \tag{9c}$$

The objective (9a) is to minimize the interval width, i.e., to maximize sharpness. Guided by this objective, preserving well reliability requires the PI coverage probability no less than the NCP. Therefore, the *soft* constraint of PI coverage defined in (9b) is introduced via probability operator $\mathbb{P}(\cdot)$ and known as *chance constraint* or *probability constraint*, which permits target misses under specific risk level β. The *hard* constraint (9c) restricts that the PI is contained in the normalized wind power range from zero to one.

III. Chance Constrained Extreme Learning Machine for Prediction Intervals

A. Extreme Learning Machine

Extreme learning machine, a single-hidden-layer feedforward neural network, has prominent advantages in terms of high nonlinear mapping capability, simple mathematical formulation, and efficient network training [34]. Suppose an ELM with I input-layer neurons, H hidden-layer neurons and O output-layer neurons, and denote the input feature vector of ELM by $\boldsymbol{x}_t \in \mathbb{R}^I$ and the hidden-layer activation function by $\varphi(\cdot)$. The hidden-layer output vector $\boldsymbol{h} \in \mathbb{R}^H$ can be expressed as

$$\boldsymbol{h}\left(\boldsymbol{x}_t;\, \{\boldsymbol{a}_i, b_i\}_{i=1}^H\right) := \begin{bmatrix} \varphi\left(\langle \boldsymbol{a}_1, \boldsymbol{x}_t\rangle + b_1\right) \\ \vdots \\ \varphi\left(\langle \boldsymbol{a}_H, \boldsymbol{x}_t\rangle + b_H\right) \end{bmatrix} \tag{10}$$

where $\boldsymbol{a}_i \in \mathbb{R}^I$ and $b_i \in \mathbb{R}$ denote the input weight vector and hidden-layer bias respectively corresponding to the ith hidden-layer neuron. Then the output vector $\boldsymbol{f} \in \mathbb{R}^O$ of ELM can be

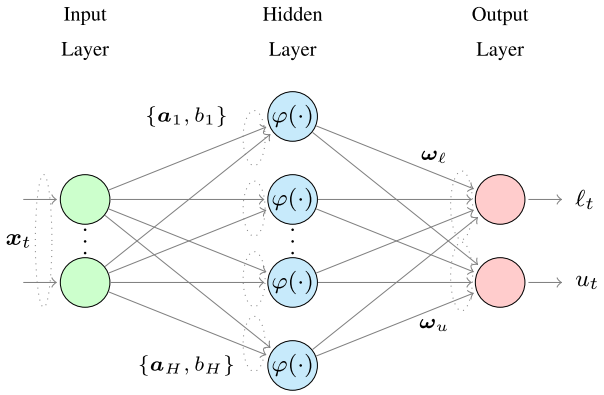


Fig. 1. Extreme learning machine for prediction interval regression.

expressed by

$$\boldsymbol{f}(\boldsymbol{x}_t) := \begin{bmatrix} f(\boldsymbol{x}_t, \boldsymbol{\omega}_1) \\ \vdots \\ f(\boldsymbol{x}_t, \boldsymbol{\omega}_O) \end{bmatrix} := \begin{bmatrix} \boldsymbol{h}\left(\boldsymbol{x}_t;\, \{\boldsymbol{a}_i, b_i\}_{i=1}^{H}\right)^\top \boldsymbol{\omega}_1 \\ \vdots \\ \boldsymbol{h}\left(\boldsymbol{x}_t;\, \{\boldsymbol{a}_i, b_i\}_{i=1}^{H}\right)^\top \boldsymbol{\omega}_O \end{bmatrix} \tag{11}$$

where $\boldsymbol{\omega}_j \in \mathbb{R}^H$ represents the output weight vector connecting all the hidden-layer neurons to the jth output-layer neuron. The input weights and hidden-layer biases $\{\boldsymbol{a}_i, b_i\}_{i=1}^H$ of ELM are randomly generated without further tuning. Therefore, only the output weight vectors $\{\boldsymbol{\omega}_i\}_{i=1}^O$ of linear system (11) remain to be optimized in training process. Fig. 1 illustrates the ELM as the regression model for producing PIs of wind power generation. The two output-layer neurons yield the lower and upper bounds of PIs, given by

$$\ell_t = f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell), \quad u_t = f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) \tag{12}$$

where $\boldsymbol{\omega}_\ell$ and $\boldsymbol{\omega}_u$ denote the output weight vectors corresponding to the lower and upper bounds ℓ_t and u_t respectively.

B. Chance Constrained Learning for Prediction Intervals

Suppose vector $\boldsymbol{x} \in \mathbb{R}^I$ and scalar $y \in \mathbb{R}$ are two random variables representing the stochasticity of input feature and target variables in sampling space respectively. Statistical learning theory generally assumes that there exists a certain *joint* probability distribution $\mu(\boldsymbol{x}, y)$, from which all sample pairs in the training dataset are drawn independently and identically [35], written as

$$(\boldsymbol{x}_t, y_t) \overset{\text{i.i.d.}}{\sim} \mu(\boldsymbol{x}, y), \ \forall t \in \mathcal{T} \tag{13}$$

where $\boldsymbol{x}_t$ and y_t are sampling realizations of input feature and target variables respectively, and $\mathcal{T}$ is the index set of training samples.

According to the ELM model for PIs (12) and the basic assumption (13) of learning theory, the decision-making framework (9) for PIs of wind power generation becomes a chance constrained extreme learning machine, given as follows

$$\min_{\boldsymbol{\omega}_\ell, \boldsymbol{\omega}_u} \quad \mathbb{E}_\mu\left[f(\boldsymbol{x}, \boldsymbol{\omega}_u) - f(\boldsymbol{x}, \boldsymbol{\omega}_\ell)\right] \tag{14a}$$

$$\text{s.t.} \quad \mathbb{P}_\mu\left(f(\boldsymbol{x}, \boldsymbol{\omega}_\ell) \le y \le f(\boldsymbol{x}, \boldsymbol{\omega}_u)\right) \ge 1 - \beta \tag{14b}$$

$$0 \le f(\boldsymbol{x}, \boldsymbol{\omega}_\ell) \le f(\boldsymbol{x}, \boldsymbol{\omega}_u) \le 1 \tag{14c}$$

Different from the chance constrained PI formulation (9), the proposed CCELM model (14) takes the joint stochasticity of the input feature and target variable pair $(\boldsymbol{x}, y)$ into consideration. However, the expectation operator $\mathbb{E}_\mu[\,\cdot\,]$ in the objective (14a) and the probability operator $\mathbb{P}_\mu(\cdot)$ in the chance constraint (14b) with respect to the probability distribution $\mu(\boldsymbol{x}, y)$ hinder the optimum attainment of CCELM model (14).

In order to avoid prior assumptions on the distribution $\mu(\boldsymbol{x}, y)$ and complex calculation of high-dimensional integral, it is common to replace the stochastic formulation (14) by the empirical one, written as

$$v^\star = \min_{\gamma_t, \boldsymbol{\omega}_\ell, \boldsymbol{\omega}_u} \sum_{t \in \mathcal{T}} \left[f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) - f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell)\right] \tag{15a}$$

$$\text{s.t.} \ \ \gamma_t = \max\left\{\begin{matrix} f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell) - y_t, \\ y_t - f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) \end{matrix}\right\}, \forall t \in \mathcal{T} \tag{15b}$$

$$\sum_{t \in \mathcal{T}} \left[1 - \mathbb{I}\left(\gamma_t \le 0\right)\right] \le \beta |\mathcal{T}| \tag{15c}$$

$$0 \le f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell) \le f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) \le 1, \forall t \in \mathcal{T} \tag{15d}$$

where $v^\star$ is the optimal value of problem (15). Objective (15a) divided by $|\mathcal{T}|$ is the estimation of expectation (14a). Positive γ_t in equality (15b) means target misses of the corresponding PIs, while negative or zero γ_t implies target hits of PIs. Constraint (15c), usually named *knapsack* constraint, budgets the counts of target misses and restricts the violation rate of target hits less than or equal to the tolerated risk level β. In the filed of chance constrained stochastic programming, such reformulation (15) is often termed *sample average approximation* [36].

Both the CCELM model (14) and its SAA counterpart (15) are free of predetermining the lower and upper proportions of PIs. Therefore, the proposed CCELM model merits sufficient adaptivity to produce central and non-central PIs with conditionally varying proportion pairs for the purpose of achieving the shortest overall interval width.

IV. Reformulation and Training Algorithm

A. Parametric Optimization and Searching Problem

The indicator function in knapsack constraint (15c) results in discontinuity of feasible region, which makes the SAA problem (15) a combinatorial optimization.

Actually, it is rather difficult to search for a feasible solution within a highly discrete region, which accounts for the complexity and hardness for attaining the optimum of combinatorial problem [37]. Convex relaxation techniques are usually employed to cope with the knapsack constraint, which tend to achieve infeasible solutions to original problem due to the weak relaxation of massive complicated indicator functions [37]. It is preferably expected to have a continuous feasible set taking the form of convex polyhedron.

To obtain a continuous feasible region, the left-hand side of knapsack constraint (15c) is moved to be the objective as the zero-one loss [38]. Meanwhile, the original objective function (15a) is combined with a synthetic parameter v to constitute an inequality constraint. The resultant problem is formulated as

follows

$$\rho(v) = \min_{\gamma_t, \boldsymbol{\omega}_\ell, \boldsymbol{\omega}_u} \sum_{t\in\mathcal{T}} [1 - \mathbb{I}(\gamma_t \le 0)] \tag{16a}$$

$$\text{s.t. } \gamma_t = \max \begin{Bmatrix} f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell) - y_t, \\ y_t - f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) \end{Bmatrix}, \forall t \in \mathcal{T} \tag{16b}$$

$$\sum_{t\in\mathcal{T}} [f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) - f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell)] \le v \tag{16c}$$

$$0 \le f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell) \le f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) \le 1, \forall t \in \mathcal{T} \tag{16d}$$

where $\rho(v)$ denotes the optimal value of the zero-one loss optimization problem (16) with the synthetic parameter being v. Since the left-hand side of constraint (16c) is the overall interval width, the synthetic parameter v can be defined within $[0, |\mathcal{T}|]$.

Actually, the zero-one loss optimization problem (16) with synthetic parameter v can be classified as *parametric optimization* (PO). Distinguished from standard optimization formulated by decision variables and constant coefficients, the PO additionally contains one or more parameters to be specified. When these parameters are specified, the PO degrades to standard optimization. In this case, the parameter v in the parametric zero-one loss optimization problem (16) reflects the budget for overall interval width and controls feasible region. Consequently, both the corresponding optimal solution and optimal value of the parametric zero-one loss optimization problem vary with the parameter v. The relationship between the parameter v and the optimal value is termed *optimal value function*.

The PZO problem (16) minimizes the target miss probability of PIs subject to the overall interval width budget v. Recall that the paradigm of probabilistic forecasting maximizes sharpness subject to well reliability [32]. This paradigm requires probing the minimum value of overall width budget v in the PZO problem (16) with the target miss probability of PIs no greater than the tolerated risk level β. Specifically, probing the minimum overall width budget corresponds to sharpness maximization, and qualifying the target miss probability corresponds to imposing reliability constraint. Therefore, the resultant parameter searching problem highly coincides with the paradigm of probabilistic forecasting.

Under mild conditions, it is claimed in *Proposition 1* and *Proposition 2* that searching for the minimum value of parameter v in the PZO (16) such that $\rho(v) \le \beta|\mathcal{T}|$ is equivalent to computing the optimal value of the SAA problem (15). For the convenience of description, the feasible regions of SAA problem (15) and the PZO problem (16) under parameter v are denoted by Ω_{SAA} and $\Omega_{\text{PO}}(v)$ respectively.

Proposition 1 (Existence of Minimum Parameter): On condition that the feasible region Ω_{SAA} of the SAA problem (15) is nonempty and the feasible region $\Omega_{\text{PO}}(v)$ of the PZO problem (16) is bounded, there exists a minimum parameter value $v = \hat{v}$ such that $\rho(v) \le \beta|\mathcal{T}|$ holds for the PZO problem (16).

Remark 1: The condition on the boundedness of feasible region for the SAA problem (15) is fairly mild. Firstly, since the wind power y_t is normalized, one can see that $|\gamma_t| \le 1$ trivially holds. Secondly, the elements of ELM output weights $\boldsymbol{\omega}_\ell$ and $\boldsymbol{\omega}_u$ cannot be infinitely large, otherwise additional constraints bounding $\|\boldsymbol{\omega}_\ell\|_1$ and $\|\boldsymbol{\omega}_u\|_1$ could be imposed.

Proposition 2 (Equivalence of Minimum Parameter): For the PZO problem (16), the minimum parameter value $v = \hat{v}$ such that $\rho(v) \le \beta|\mathcal{T}|$ equals the optimal value $v^\star$ of the original SAA problem (15).

The pointwise maximum function in the constraint (16b) makes the feasible region $\Omega_{\text{PO}}(v)$ nonconvex. However, as shown by *Proposition 3*, the equality constraint (16b) can be exactly relaxed into inequalities. Then the original PZO problem (16) is equivalently transformed into the relaxed form, given by

$$\underline{\rho}(v) = \min_{\gamma_t, \boldsymbol{\omega}_\ell, \boldsymbol{\omega}_u} \sum_{t\in\mathcal{T}} [1 - \mathbb{I}(\gamma_t \le 0)] \tag{17a}$$

$$\text{s.t. } \gamma_t \ge f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell) - y_t, \forall t \in \mathcal{T} \tag{17b}$$

$$\gamma_t \ge y_t - f(\boldsymbol{x}_t, \boldsymbol{\omega}_u), \forall t \in \mathcal{T} \tag{17c}$$

$$\sum_{t\in\mathcal{T}} [f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) - f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell)] \le v \tag{17d}$$

$$0 \le f(\boldsymbol{x}_t, \boldsymbol{\omega}_\ell) \le f(\boldsymbol{x}_t, \boldsymbol{\omega}_u) \le 1, \forall t \in \mathcal{T} \tag{17e}$$

The feasible region determined by the constraints (17b)–(17e) is denoted by $\overline{\Omega}_{\text{PO}}(v)$. By virtue of linear formulation of ELM, the resultant feasible region $\overline{\Omega}_{\text{PO}}(v)$ is essentially a convex polyhedron.

Proposition 3 (Exactness of Inequality Relaxation): The inequality relaxation (17b)–(17c) of the equality constraint (16b) does not change the optimal value of the original PZO problem (16), namely, $\rho(v) = \underline{\rho}(v)$.

The proofs of *Propositions 1–3* are collected in Appendix.

Since the feasible region of the relaxed PZO problem (17) gets enlarged as the overall width budget v increases, the optimal value function $\underline{\rho}(v)$ is monotonically non-increasing. If the optimal value function $\underline{\rho}(v)$ at a certain parameter value is less than or equal to the threshold $\beta|\mathcal{T}|$, this parameter value could be further reduced. Otherwise, the parameter value should be increased to achieve the reliability requirement $\underline{\rho}(v) \le \beta|\mathcal{T}|$. Thus, given an oracle capable of evaluating the optimal value function $\underline{\rho}(v)$, bisection search can be utilized to obtain the minimum parameter value $v = \underline{v}$ such that $\underline{\rho}(v) \le \beta|\mathcal{T}|$ with the absolute error no greater than a preset tolerance. The minimum parameter value that maintains the optimal value of PO problems no greater than $\beta|\mathcal{T}|$ is hereafter referred to as *goal parameter value* for the corresponding parameter searching tasks.

Remark 2: Proposition 1 and *Proposition 2* reveal that the SAA problem (15) with discrete feasible region Ω_{SAA} can be reformulated as a parameter searching task in the PZO problem (16) with continuous and nonsmooth feasible region $\Omega_{\text{PO}}(v)$. *Proposition 3* shows that such continuous and nonsmooth feasible region can be equivalently transformed into a convex polyhedron $\overline{\Omega}_{\text{PO}}(v)$, which allows attaining a feasible solution easily.

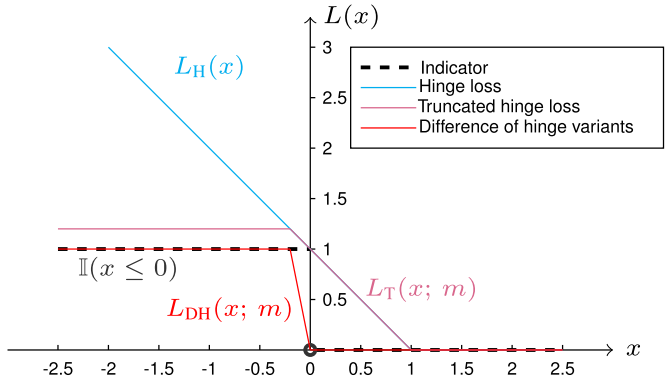


Fig. 2. Graphs of different surrogates for the indicator function ($m = 5$).

B. Surrogates for Indicator Function

However, directly minimizing the zero-one loss (17a) is also a combinatorial optimization problem and known to be $\mathcal{NP}$-complete [39]. Therefore, a surrogate for the indicator function $\mathbb{I}(x \le 0)$ involved in the zero-one loss (17a) is indispensable to achieve computational tractability.

1) Hinge Loss: Hinge loss is widely applied to classification problem, formulated as the following convex function, given by

$$L_{\mathrm{H}}(x) := \max\{1 - x,\, 0\} \tag{18}$$

2) Truncated Hinge Loss: Since the hinge loss $L_{\mathrm{H}}(x)$ is sensitive to negative outliers, the truncated hinge loss is designed in [38], defined as

$$L_{\mathrm{T}}(x;\, m) := L_{\mathrm{H}}(x) - \max\left\{-\tfrac{1}{m} - x,\, 0\right\} \tag{19}$$

where the positive parameter m determines the truncation point.

3) Difference of Hinge Variants: Inspired by the previous loss functions, a new surrogate for indicator function with high flexibility and accuracy is elaborated in the study, which is formulated by the difference of hinge loss variants

$$\begin{aligned} L_{\mathrm{DH}}(x;\, m) :\ &= \max\{-mx,\, 0\} - \max\{-mx - 1,\, 0\} \\ &= \begin{cases} \mathbb{I}(x \le 0) & \text{if } x \in \left(-\infty,\, -\frac{1}{m}\right] \cup (0, +\infty) \\ -mx & \text{if } x \in \left(-\frac{1}{m},\, 0\right] \end{cases} \end{aligned} \tag{20}$$

where $m > 0$ is the slope parameter controlling the accuracy. The larger the parameter m is, the higher the resemblance degree between the surrogate and indicator is.

Remark 3: The elaborated difference of hinge variants (20) can be categorized as difference of convex functions. It serves as an underestimator of the indicator if $x \in (-\frac{1}{m},\, 0]$, and exactly equals the indicator in other parts of the domain.

The aforementioned surrogates for the indicator function are visually shown in Fig. 2. Apparently, the proposed difference of hinge variants (20) resembles the indicator function best.

C. DC Optimization

By substituting the difference of hinge variants (20) for the indicator in objective function (17a) and collecting the convex terms, the objective function (17a) is approximated by

$$L(\boldsymbol{\gamma}) := \sum_{t \in \mathcal{T}} [1 - L_{\mathrm{DH}}(\gamma_t)] := L^{+}_{\mathrm{vex}}(\boldsymbol{\gamma}) - L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma}) = \underbrace{\sum_{t \in \mathcal{T}} [1 + \max\{-m\gamma_t - 1,\, 0\}]}_{L^{+}_{\mathrm{vex}}(\boldsymbol{\gamma})} - \underbrace{\sum_{t \in \mathcal{T}} [\max\{-m\gamma_t,\, 0\}]}_{L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma})} \tag{21}$$

The objective approximation $L(\boldsymbol{\gamma})$ is indeed an upper bound of the zero-loss summation (17a) in the form of difference of convex functions, where $\boldsymbol{\gamma} := [\gamma_1\ \gamma_2\ \cdots\ \gamma_{|\mathcal{T}|}]^{\top}$ is the vector collecting γ_t for all $t \in \mathcal{T}$. Then the resultant parametric DC optimization problem is formulated as

$$\overline{\rho}(v) = \min_{\boldsymbol{\theta}} \quad L(\boldsymbol{\gamma}) \tag{22a}$$

$$\text{s.t.} \quad \boldsymbol{\theta} \in \overline{\Omega}_{\mathrm{PO}}(v) \tag{22b}$$

where $\overline{\rho}(v)$ is the optimal value function of parametric DC optimization (22), and $\boldsymbol{\theta} := [\boldsymbol{\gamma}^{\top}\ \boldsymbol{\omega}_{\ell}^{\top}\ \boldsymbol{\omega}_{u}^{\top}]^{\top}$ is the vector concatenating all decision variables.

Problem (22) minimizes the difference of convex functions $L^{+}_{\mathrm{vex}}(\boldsymbol{\gamma})$ and $L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma})$ subject to the polyhedral feasible region $\overline{\Omega}_{\mathrm{PO}}(v)$ defined by inequalities (17b)–(17e), which can be solved via the DC optimization algorithm (DCA) presented in Algorithm 1. The DCA only requires tackling sequential convex optimization subproblems and guarantees to converge to a high-quality optimum in finite iterations [40]. The key procedures are explained as follows.

1) Iterative Formulation: In general, the DCA replaces the convex subtrahend term of objective function by its global affine underestimator and obtains a convex optimization problem. For the DC optimization problem (22), the solution of the $(k+1)$th iteration is updated according to

$$\boldsymbol{\theta}^{(k+1)} = \arg \min_{\boldsymbol{\theta} \in \overline{\Omega}_{\mathrm{PO}}(v)} L^{+}_{\mathrm{vex}}(\boldsymbol{\gamma}) - \underbrace{[L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma}^{(k)}) + \boldsymbol{\delta}^{(k)\top}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(k)})]}_{\text{global affine underestimator of } L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma})} \tag{23}$$

where $\boldsymbol{\delta}^{(k)}$ is a subgradient of $L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma}^{(k)}$, defined as

$$\boldsymbol{\delta}^{(k)} \in \left\{ \boldsymbol{g} \in \mathbb{R}^{|\mathcal{T}|} \mid L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma}) \ge L^{-}_{\mathrm{vex}}(\boldsymbol{\gamma}^{(k)}) + \boldsymbol{g}^{\top}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(k)}), \forall \boldsymbol{\gamma} \right\}$$

Algorithm 1: DC Optimization Routine: $\mathtt{DCA}(\epsilon, m, v, \boldsymbol{\theta}_0)$.

Input:
tolerance ϵ;
slope m of surrogate function;
overall interval width budget v;
initial solution $\boldsymbol{\theta}_0$;
Output:
converged optimum $\overline{\boldsymbol{\theta}}$;
1: $k \leftarrow 0$; $\boldsymbol{\theta}^{(k)} \leftarrow \boldsymbol{\theta}_0$; $\|\boldsymbol{e}\|_2 \leftarrow +\infty$
2: **while** $\|\boldsymbol{e}\|_2 \ge \epsilon$ **do**
3: Obtain $\boldsymbol{\theta}^{(k+1)}$ by solving LP (25);
4: $k \leftarrow k + 1$;
5: $\boldsymbol{e} \leftarrow \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}$;
6: **end while**
7: $\overline{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(k)}$;

$$= \left\{ \left[g_1\ g_2\ \cdots\ g_{|\mathcal{T}|}\right]^\top \middle| \begin{array}{l} \forall t \in \mathcal{T}, \\ \begin{cases} g_t = -m & \text{if } \gamma_t^{(k)} < 0 \\ g_t \in [-m, 0] & \text{if } \gamma_t^{(k)} = 0 \\ g_t = 0 & \text{if } \gamma_t^{(k)} > 0 \end{cases} \end{array} \right\} \tag{24}$$

Without loss of generality, this paper sets $g_t = 0$ if $\gamma_t^{(k)} = 0$.

2) Nonsmooth Objective Linearization: By minimizing a convex function over a polyhedron, problem (23) is a convex optimization problem. With the constant items $L_{\text{vex}}^{-}(\boldsymbol{\gamma}^{(k)})$ and ${\boldsymbol{\delta}^{(k)}}^\top \boldsymbol{\gamma}^{(k)}$ omitted, problem (23) with nonsmooth objective can be equivalently linearized into the epigraphical problem form, expressed as

$$\min_{\boldsymbol{\tau},\,\boldsymbol{\theta}} \quad \mathbf{1}^\top \boldsymbol{\tau} - {\boldsymbol{\delta}^{(k)}}^\top \boldsymbol{\gamma} \tag{25a}$$

$$\text{s.t.} \quad \boldsymbol{\tau} \geq \mathbf{1},\ \boldsymbol{\tau} \geq -m\boldsymbol{\gamma} \tag{25b}$$

$$\boldsymbol{\theta} \in \overline{\Omega}_{\text{PO}}(v) \tag{25c}$$

where $\boldsymbol{\tau} \in \mathbb{R}^{|\mathcal{T}|}$ is the introduced auxiliary vector; $\mathbf{0}$ and $\mathbf{1}$ denote the $|\mathcal{T}|$-dimensional vectors with all the elements being zero and one respectively. It can be easily seen that the problem (25) is a LP model, which can be solved effectively and efficiently.

3) Initial Solution Determination: The iterative formulation (23) requires an initial feasible solution $\boldsymbol{\theta}_0 \in \overline{\Omega}_{\text{PO}}(v)$ in the first iteration. The LP problem that minimizes the overall target deviation from the nearest PI bound is solved to provide an initial solution $\boldsymbol{\theta}_0$, given by

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \left\{ \mathbf{1}^\top \boldsymbol{\gamma} \mid \boldsymbol{\gamma} \geq \mathbf{0},\ \boldsymbol{\theta} \in \overline{\Omega}_{\text{PO}}(v) \right\} \tag{26}$$

In the initial solution determination problem (26), the target deviation γ_t is defined as the minimum distance from the missed target y_t to the corresponding lower and upper bounds of PI. The deviation would be zero if the target y_t successfully hits the corresponding PI.

D. DC Optimization Based Bisection Search

As analyzed in Subsection IV-A, training the CCELM model suffices to search for the goal parameter value $\underline{v}$ in the PZO problem (17). Bisection search can be leveraged to obtain the goal parameter value given the monotonicity of the optimal function value $\underline{\rho}(v)$. Instead of dealing with the PZO problem (17) directly, the parameter searching is performed with respect to its DC approximation form (22) due to the computational tractability. Note that the zero-one loss optimization problem (17) and the DC optimization problem (22) minimize similar objective functions subject to common feasible region $\overline{\Omega}_{\text{PO}}(v)$. The optimum of the DC optimization problem (22) can be a high-quality approximate solution to the zero-one loss minimization problem (17), since both solutions achieve close objective values.

Given a certain parameter v and certain initial solution $\boldsymbol{\theta}_0$, a local optimum $\overline{\boldsymbol{\theta}}$ of the DC optimization problem (22) can be obtained by the routine $\texttt{DCA}(\epsilon, m, v, \boldsymbol{\theta}_0)$ in Algorithm 1. By evaluating whether the objective function value attained at the optimum $\overline{\boldsymbol{\theta}}$ is greater than the threshold $\beta|\mathcal{T}|$, forecasters are able to estimate whether the overall interval width budget v could be further reduced without the empirical probability of target misses exceeding the tolerated risk level β.

Algorithm 2: DC Optimization based Bisection Search.

Input:
outer loop tolerance ϵ_1 for bisection search;
inner loop tolerance ϵ_2 for DCA;
slope m of surrogate function;
lower and upper bounds $\underline{v}_\ell$ and $\underline{v}_u$ for the goal parameter value;
Output:
converged overall interval width budget $\overline{v}$;
output weights of ELM $\boldsymbol{\omega}_\ell$ and $\boldsymbol{\omega}_u$;

1: **while** $\underline{v}_u - \underline{v}_\ell > \epsilon_1$ **do**
2: $\overline{v} \leftarrow (\underline{v}_\ell + \underline{v}_u)/2$;
3: Set $\boldsymbol{\theta}_0$ as the minimizer of (26);
4: Query DC optimization routine in Algorithm 1: $\overline{\boldsymbol{\theta}} \leftarrow \texttt{DCA}(\epsilon_2, m, \overline{v}, \boldsymbol{\theta}_0)$;
5: Extract $\overline{\boldsymbol{\omega}}_\ell$ and $\overline{\boldsymbol{\omega}}_u$ from $\overline{\boldsymbol{\theta}}$;
6: Calculate the target miss counts of PIs

$$\texttt{miss} \leftarrow \sum_{t \in \mathcal{T}} \left[1 - \mathbb{I}\left(f(\boldsymbol{x}_t, \overline{\boldsymbol{\omega}}_\ell) \leq y_t \leq f(\boldsymbol{x}_t, \overline{\boldsymbol{\omega}}_u)\right)\right];$$

7: **if** $\texttt{miss} \leq \beta|\mathcal{T}|$ **then**
8: $\underline{v}_u \leftarrow \overline{v}$; $\boldsymbol{\omega}_\ell \leftarrow \overline{\boldsymbol{\omega}}_\ell$; $\boldsymbol{\omega}_u \leftarrow \overline{\boldsymbol{\omega}}_u$;
9: **else**
10: $\underline{v}_\ell \leftarrow \overline{v}$;
11: **end if**
12: **end while**

Consequently, the DC optimization based bisection search algorithm is proposed in Algorithm 2 to approximately achieve the goal parameter value $\underline{v}$ and efficiently fulfill model training. Given the initial searching range $[\underline{v}_\ell, \underline{v}_u]$ enclosing the goal parameter value $\underline{v}$ for sure, the midpoint $(\underline{v}_\ell + \underline{v}_u)/2$ serves as the incumbent parameter value $\overline{v}$ to be evaluated (line 2). The solution obtained by DC optimization routine under the incumbent parameter value $\overline{v}$ (lines 3 to 4) approximates the optimum of the zero-one loss minimization problem (17) from above. Then the value of target miss counts is calculated as an overestimation of the corresponding optimal value $\underline{\rho}(\overline{v})$ (lines 5 to 6). If the target miss counts achieved at this minimum are less than or equal to the threshold $\beta|\mathcal{T}|$, the goal parameter $\underline{v}$ must exist in $[\underline{v}_\ell, \overline{v}]$ (lines 7 to 8) due to the non-increasing monotonicity of the optimal value function $\underline{\rho}(v)$. Otherwise, the lower bound $\underline{v}_\ell$ of searching range is approximately set to be the incumbent parameter $\overline{v}$ (lines 9 to 10). The iteration terminates when the preset tolerance ϵ_1 is reached, which means that the goal parameter $\underline{v}$ is expected to stay in the range $[\underline{v}_\ell, \underline{v}_u]$ with width less than ϵ_1.

Remark 4: The output weight vectors $\boldsymbol{\omega}_\ell$ and $\boldsymbol{\omega}_u$ of ELM returned by Algorithm 2 are feasible to the SAA problem (15) on condition that the counts of target misses do not exceed $\beta|\mathcal{T}|$ under the returned parameter value $\overline{v}$. Thus, the returned parameter value $\overline{v}$ can be regarded as an upper approximation of the goal parameter value $\underline{v}$. Although the exact minimum overall interval width $\underline{v}$ is extremely hard to obtain in practice, the upper approximation $\overline{v}$ provided by Algorithm 2 is sufficient to preserve the superior sharpness performance of the CCELM based proportion-free PIs.

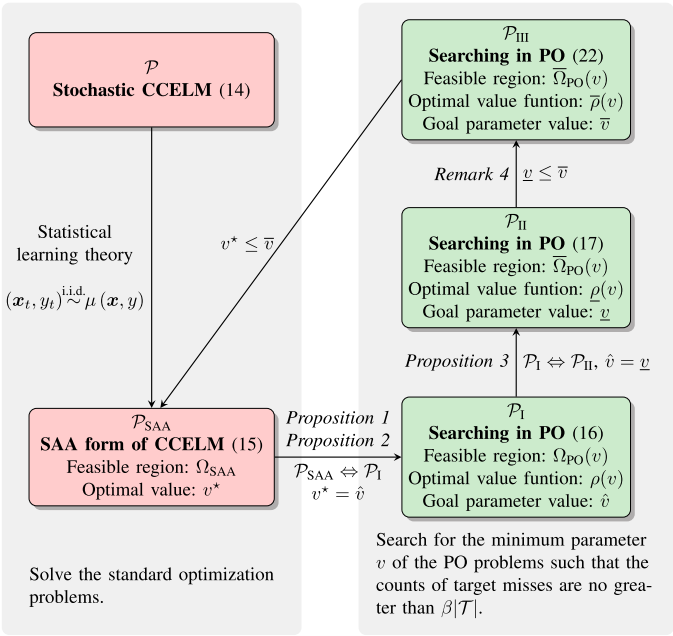


Fig. 3. Schematic framework of CCELM model transformations.

Schematic framework of the aforementioned CCELM model transformations is summarized in Fig. 3, including the standard optimization formulations $\mathcal{P}$ and $\mathcal{P}_{\text{SAA}}$, as well as the parameter searching formulations $\mathcal{P}_{\text{I}}$, $\mathcal{P}_{\text{II}}$, and $\mathcal{P}_{\text{III}}$. Actually, the parameter searching problem $\mathcal{P}_{\text{III}}$ solved via Algorithm 2 acts as a proxy for parameter searching formulations $\mathcal{P}_{\text{I}}$ and $\mathcal{P}_{\text{II}}$, both of which are equivalent to solving the SAA optimization problem $\mathcal{P}_{\text{SAA}}$. Note that Algorithm 2 successfully accomplishes a machine learning task by solving a finite sequence of convex LP problems. Therefore, the proposed DCBS algorithm merits remarkable efficiency and practicality.

V. Numerical Studies

A. Description of Experiments

Actual wind power data from Glens of Foudland wind farm [41] is applied to the numerical experiments. Glens of Foudland wind farm, located in the county of Aberdeenshire, Scotland, U.K., has twenty wind turbines with the total nominal capacity of 26 MW. The wind power series with 30-minute resolution in 2017 are utilized in this study, which are normalized with respect to the nominal capacity. In each season, about 60% of the wind power data builds the training dataset, and the rest serves as the test dataset.

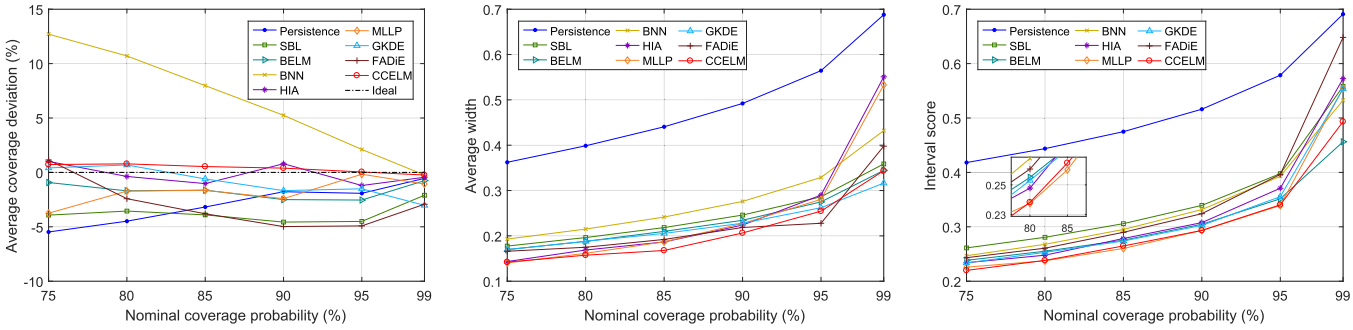
To validate the effectiveness and superiority of the proposed CCELM model, eight parametric and nonparametric probabilistic forecasting methods are used as benchmarks, including persistence, SBL, BELM, BNN, HIA, MLLP, GKDE, and FADiE. As a parametric benchmark, the persistence perceives the Gaussian probability distribution of wind power with mean being the last observation and variance calculated by the latest observations available. The SBL method generates conditional Gaussian probability distribution based on the relevance vector machine and Bayesian inference [21]. The BELM [12] and BNN [10], [11] establish a collection of mean and variance regression models in terms of the bootstrap datasets, which adopt the censored and original Gaussian assumption on the wind power forecasting uncertainty respectively. Four rival nonparametric approaches are employed here. The HIA method optimizes the interval score (8) with the absolute ACD as penalization [18]. The MLLP approach produces quantiles with specific proportion pair determined by sensitivity analysis and constructs nonparametric PIs [30]. Equipped with Gaussian and diffusion kernels respectively, the GKDE and FADiE methods firstly estimate the density function of each training target and then derive the PIs to construct the training database learned by ELM [22].

To comprehensively evaluate various forecasting approaches, central PIs generated by the aforementioned benchmarks are compared with the CCELM based PIs considering multiple confidence levels, look-ahead steps, and seasonal patterns. In addition to the central cases, the MLLP method constructs both central and non-central PIs by conducting sensitivity analysis on multiple PMBs. The resultant PIs with different proportion pairs are then compared with the CCELM based proportion-free PIs. Moreover, the computational efficiency and training algorithm performance are also analyzed.

B. Comparison of Forecasting Results

1) Multi-Confidence Prediction Intervals: Since only high-confidence PIs are of meaningful practicality to power systems, the NCPs ranging from 75% to 99% are focused in this study [42]. Without loss of generality, multi-confidence PIs with 1-hour look-ahead time in summer are constructed. The corresponding reliability, sharpness, and overall skill diagrams are depicted in Fig. 4(a), 4(b), and 4(c) respectively.

According to the reliability diagram in Fig. 4(a) and sharpness diagram in Fig. 4(b), persistence based PIs fail to preserve satisfactory reliability under NCPs from 75% to 85%, though with double interval width of other models. The comparatively large absolute ACDs over 4.5% with 90% and 95% NCPs for the parametric SBL method verify the non-Gaussian properties of prediction errors. Actually, parametric Gaussian assumption can result in severe reliability decay especially for the intermediate confidence levels. For example, ECPs of the BNN approach overshoot the 75% and 80% NCPs by more than 10%. Such uncertainty overestimation could be greatly relieved using censored Gaussian distribution. In light of the bounded property of wind power generation, the BELM method truncates the probability masses of original Gaussian distribution lying outside the unit interval. As a consequence, the ACDs of BELM model shown in Fig. 4(a) are less than 2.54% and maintain acceptable reliability. The FADiE based PIs endure absolute ACD deterioration reaching 4.8%, which can be attributed to the adopted unconditional density estimations. The nonparametric HIA, MLLP, and GKDE methods preserve relatively high reliability under the NCPs from 80% to 95%. The proposed CCELM model maintains superior reliability consistency under all the investigated confidence levels, where the largest ACD reaches only 0.79%. Such reliability fidelity to the NCP strongly validates the effectiveness of the adoption of reliability constraint



(a) Reliability diagram (b) Sharpness diagram (c) Overall skill diagram

Fig. 4. Quality comparison of multi-confidence PIs in summer with 1-hour look-ahead time: (a) reliability diagram, (b) sharpness diagram, (c) overall skill diagram.

in the CCELM model. Moreover, the proposed CCELM model merits the best sharpness performance under the NCPs from 80% to 90%. As for the cases of 75%, 95%, and 99% NCPs, the CCELM model achieves the shortest interval width among all the reliable forecasting models that maintain the ACDs within $\pm 2\%$. For example, the CCELM generates at least 10% shorter PIs than all the applied benchmarks under 85% NCP owing to the objective for minimizing the interval width. To summarize, directly setting well reliability as the constraint and sharpness maximization as the objective in the proposed CCELM model contributes to the shorter PI width and reliable ECP respectively.

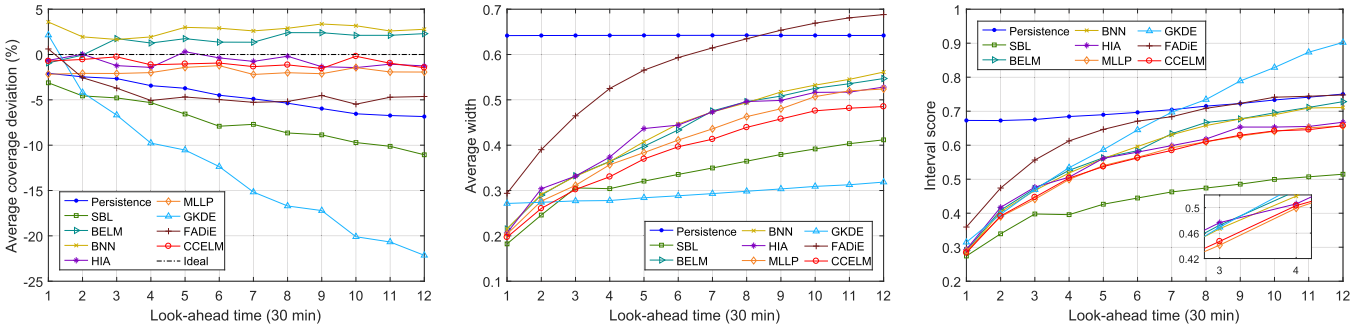
As shown in Fig. 4(c), the higher interval score S_β of the persistence, SBL, and BNN indicates the inferior overall skill of these parametric methods to the nonparametric benchmarks such as the HIA, MLLP, GKDE, and FADiE under most of the confidence levels. The interval score S_β highly appraises MLLP and CCELM approaches under different NCPs, in accordance with their excellent performance in both reliability and sharpness. It should be highlighted that the MLLP method produces PIs with inferior reliability and sharpness to the CCELM under all the investigated confidence levels, while the interval score S_β in the magnification box of Fig. 4(c) gives an incorrect clue. Higher interval score is attained for the CCELM method in comparison with the MLLP under 80% and 85% NCPs. Although the BELM approach achieves the lowest interval score S_β under 99% NCP, its sharpness performance does not demonstrate advantage over the CCELM and the associated ECP deviation is larger than the CCELM based PIs. Actually, the proposed CCELM method is independent of proportion predetermination, and the corresponding PIs can be non-central. Since the interval score S_β is derived based on the premise of central PIs [32], it naturally has preference for the MLLP and BELM based central PIs.

To conclude, the proposed CCELM model is credit with trustworthy reliability and sharpness under various confidence levels, which can be dependably put into practice with different reliability and risk requirements.

2) Multi-Step Prediction Intervals: Short-term wind power forecasting with look-ahead time up to 6 hours is essential for economic dispatch, electricity trading, and congestion management in power system operations [43]. To justify the forecasting application feasibility with respect to a wide range of time horizons, PIs with look-ahead time from 30 to 360 minutes are generated, under which utilizing only historical wind power measurements as the input features is sufficient to build accurate forecasting models [20]. Quality assessment criteria of multi-step PIs under 90% NCP in spring are graphically illustrated in Fig. 5.

It can be found from Fig. 5(a) that the persistence, SBL, and GKDE methods suffer from reliability degradation to some extent as the look-ahead time increases. Under 1-hour look-ahead time, the absolute ACDs of the SBL and GKDE based PIs are larger than 4%, which might lack sufficient accuracy to fit the primary reliability requirement of forecasting. The reliability deterioration is particularly significant under the GKDE's curve, which demonstrates the deficiency of Gaussian kernel in representing the predictive density distributions under longer lead time. Although the BNN based PIs sustain stable reliability performance under different look-ahead steps, the resultant positive ACDs from the 90% NCP exceed 2.6% in most of the cases. In contrast, the BELM, HIA, MLLP, and CCELM methods have promising and consistent reliability performance, with the absolute ACDs less than 2.41%.

The average interval width comparisons are illustrated in the sharpness diagram Fig. 5(b). Intuitively, there would be more uncertainties involved in future wind power generation under longer lead time, and the AW of PIs usually increases with the look-ahead steps. However, the nearly constant AW of the persistence based PIs reflects the lack of situational adaptivity. The FADiE method tends to overreact to the increasing future uncertainties, and the resulting predictive PDFs are relatively dispersive. This uncertainty overreaction of the FADiE approach leads to more than 40% wider PIs than the CCELM within the investigated six-hour look-ahead steps. The SBL and GKDE slightly increase their PI width with the lead time and cannot preserve satisfactory reliability performances under longer look-ahead horizons. In general, the AWs achieved by the BNN approach are a bit wider than those of the BELM based PIs, resulted from the unbounded Gaussian assumption of BNN and the superior generalization capability of ELM. It is worth noting that the BELM, HIA and MLLP methods generate reliable PIs in comparison with the CCELM model. Nevertheless, the proposed CCELM model consistently achieves higher sharpness than these competitively reliable benchmarks. The potentially asymmetric PDFs of wind power forecasting errors under longer



(a) Reliability diagram (b) Sharpness diagram (c) Overall skill diagram

Fig. 5. Quality comparison of multi-step PIs in spring with 90% NCP: (a) reliability diagram, (b) sharpness diagram, (c) overall skill diagram.

lead time are generally less concentrated and have heavier tails [42], of which the central PIs significantly suffer from interval width conservativeness. Therefore, the sharpness superiority of CCELM model becomes more apparent with the increase of look-ahead time. The AW of the CCELM based PIs with 6-hour look-head time reaches 0.4852, and is 13.7%, 8.1% and 7.4% shorter than that of the BELM, HIA, and MLLP respectively, verifying the advantages of adaptive proportion-free PIs in terms of width shortening.

The interval score values under various look-ahead steps are also provided in Fig. 5(c). Derived from predictive Gaussian distributions, the SBL based central PIs are remarkably favored by the interval score S_β. However, these PIs have inappropriate sharpness and unqualified ACDs dropping to about -10% under 5-hour look-ahead time, which indicates that the interval score is lack of capability in terms of discriminating the contributions from reliability and sharpness [44]. Despite that the interval score S_β for the CCELM based PIs with 90-minute look-ahead time is larger than that of the MLLP based PIs, both reliability and sharpness diagrams uphold the dominant advantages of the CCELM method. The overall skill misjudgement in Fig. 5(c) implies the interval score's unsuitability for evaluating PIs free of bounds proportions.

3) Analysis of Seasonality: With consideration of seasonality, wind power datasets under different seasons are utilized for further verifications of forecasting models. Detailed forecasting performance metrics of PIs with 95% NCP and 30-minute look-ahead time are provided in Table I.

It can be observed from Table I that all the employed forecasting methods obtain significantly wider PIs in the cases of winter and autumn. Since extreme meteorological conditions usually lead to more uncertainties involved in wind power generation, the less concentrated wind power PIs in the winter and autumn cases imply the severe atmospheric changes happening during these time spans. Throughout the four seasons, the proposed CCELM model generates considerably reliable PIs with absolute ACDs less than 1.5% in comparison with the adopted benchmarks. Meanwhile, the CCELM based PIs achieve remarkable sharpness advantage even compared with the well-performed HIA and MLLP methods. Moreover, the PIs obtained by the GKDE and FADiE methods in the four seasons are at least 16% wider than those obtained by the proposed CCELM model. It is noteworthy that both the reliability and sharpness criteria in the summer and autumn cases substantially support the superiority of the proposed CCELM method to the MLLP. However, the negatively oriented interval score S_β of the CCELM is larger than that of the MLLP, which again verifies the interval score's prejudice against the CCELM based PIs. Similar prejudice can also be indicated by the BELM and CCELM based PIs. In the autumn case, the CCELM gains 5.2% higher sharpness

TABLE I
QUALITY OF PIS WITH 95% NCP AND 30-MINUTE LOOK-AHEAD TIME UNDER DIFFERENT SEASONS

Season	Method	ECP	ACD	AW	S_β
Winter	Persistence	95.18%	0.18%	0.8723	0.8852
	SBL	90.92%	-4.08%	0.2971	0.5213
	BELM	92.80%	-2.20%	0.3081	0.4447
	BNN	94.79%	-0.21%	0.3442	0.4542
	HIA	93.19%	-1.81%	0.3408	0.4847
	MLLP	93.66%	-1.34%	0.3332	0.4873
	GKDE	92.05%	-2.95%	0.3808	0.5016
	FADiE	92.90%	-2.10%	0.3944	0.5100
	CCELM	93.66%	-1.34%	0.3268	0.4610
Spring	Persistence	93.20%	-1.80%	0.7334	0.7495
	SBL	93.58%	-1.42%	0.2714	0.3820
	BELM	93.86%	-1.14%	0.2442	0.3449
	BNN	96.07%	1.07%	0.2586	0.3517
	HIA	95.40%	0.40%	0.2648	0.3559
	MLLP	93.96%	-1.04%	0.2619	0.3506
	GKDE	93.39%	-1.61%	0.3110	0.3737
	FADiE	92.72%	-2.28%	0.3148	0.4145
	CCELM	94.06%	-0.94%	0.2423	0.3376
Summer	Persistence	92.98%	-2.02%	0.5643	0.5779
	SBL	91.53%	-3.47%	0.2128	0.3392
	BELM	92.98%	-2.02%	0.2074	0.3142
	BNN	96.38%	1.38%	0.2253	0.3102
	HIA	95.45%	0.45%	0.2059	0.3021
	MLLP	93.08%	-1.92%	0.2042	0.2792
	GKDE	95.76%	0.76%	0.2621	0.3110
	FADiE	94.01%	-0.99%	0.2704	0.3362
	CCELM	94.42%	-0.58%	0.1952	0.2846
Autumn	Persistence	95.64%	0.64%	0.8426	0.8591
	SBL	92.38%	-2.62%	0.3572	0.4866
	BELM	95.10%	0.10%	0.3895	0.4664
	BNN	97.28%	2.28%	0.4103	0.4818
	HIA	96.28%	1.28%	0.3928	0.4769
	MLLP	94.37%	-0.63%	0.3947	0.4651
	GKDE	96.37%	1.37%	0.4294	0.4996
	FADiE	93.10%	-1.90%	0.4482	0.5453
	CCELM	95.01%	0.01%	0.3693	0.4687

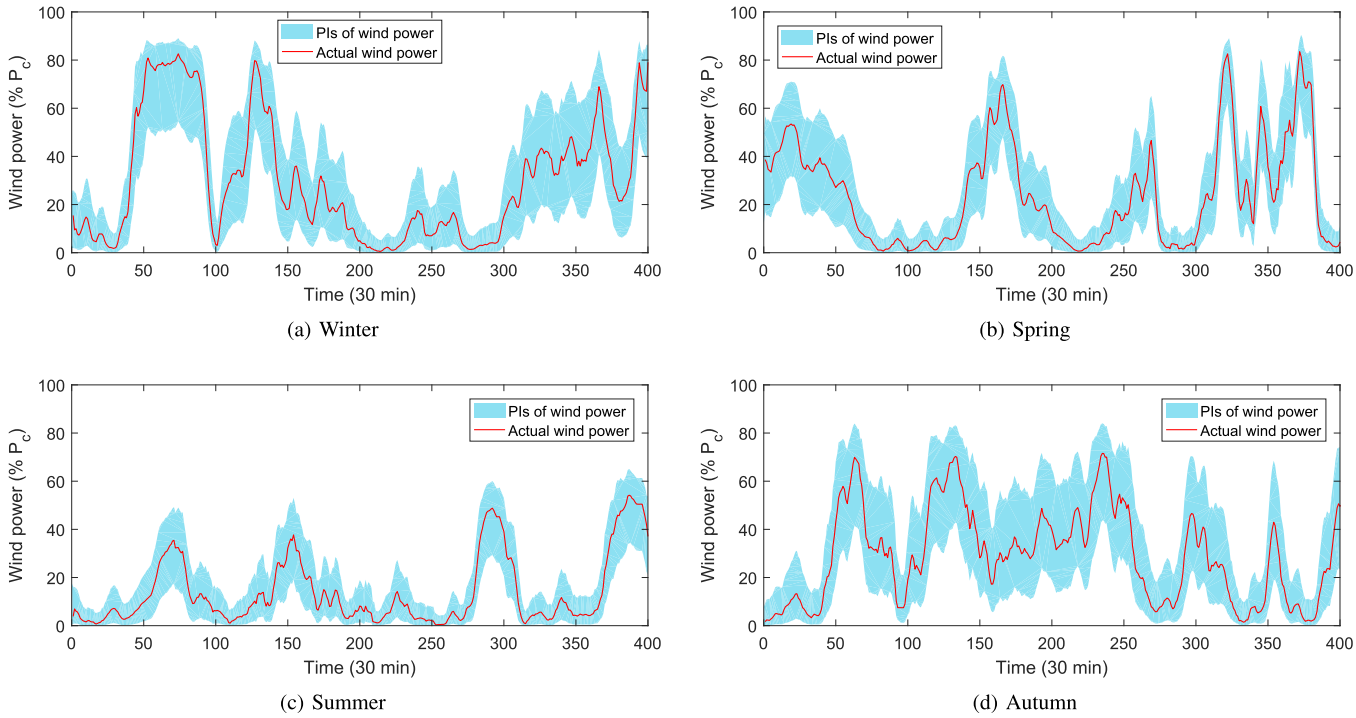


Fig. 6. PIs with 95% NCP and 30-minute look-ahead time obtained by the proposed CCELM model: (a) winter, (b) spring, (c) summer, (d) autumn.

than the BELM and superior reliability performance. But the interval score S_β makes misleading judgement that contradicts the dominant overall skill superiority of CCELM based PIs.

In order to visually demonstrate the PI quality and seasonal patterns of data, the CCELM based PIs with 95% NCP and 30-minute look-ahead time in the four seasons are depicted in Fig. 6. As clearly displayed in Fig. 6(a) and 6(d), the wind power undergoes steep ramp events in the winter and autumn cases. It can be found that the PIs corresponding to the low power levels are remarkably narrower than the medium and high power levels, which reveals more uncertainties involved in the medium and high levels of wind power generation. Actually, quality PIs should make conditional estimations of wind power prediction uncertainty. In this context, the CCELM model can perceive different uncertainty situations conveyed by the input features and generate PIs with significant width variability, which substantially indicates its excellent conditional adaptivity.

Based on the forecasting performance under different seasons, it can be concluded that the proposed CCELM method is considerably adaptive to the seasonal variations and maintains consistent superiority.

4) Advantage of Proportion-Free Prediction Intervals: For the MLLP model [30], central and non-central PIs with specific proportion pairs can be constructed via conducting sensitivity analysis on the PMB defined by (4). The proposed CCELM model is characterized by its independence of proportion predetermination. To manifest the benefits from such proportion independency, the quality of CCELM based PIs is compared with that of the MLLP based central and non-central PIs.

As a representative example, Fig. 7 displays the resultant reliability and sharpness metrics in the autumn case with 90% confidence level and 1-hour look-ahead time. The sensitivity analysis

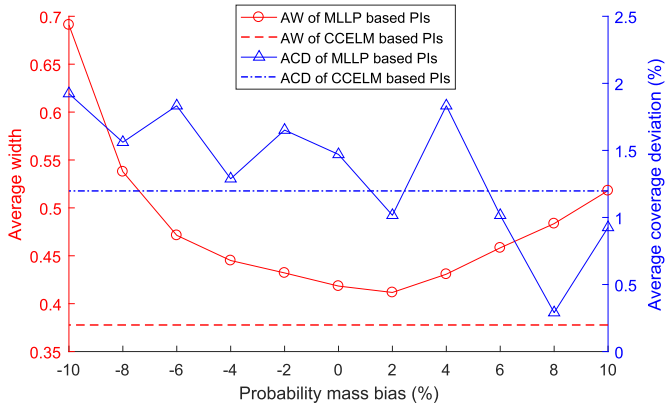


Fig. 7. Quality comparison of the MLLP based proportion-fixed PIs and the CCELM based proportion-free PIs.

of the MLLP based PIs covers the PMBs ranging from -10% to 10% with 2% resolution. It is depicted that both the MLLP and CCELM based PIs merit satisfactory reliability performance with ACDs ranging from 0.9% to 1.9%, whereas there exists significant sharpness discrepancy between the proportion-fixed and proportion-free PIs. For the MLLP based PIs with different PMBs, the AW sharply drops from the peak value 0.69 at -10% PMB to the minimum value 0.41 at 2% PMB. The 2% PMB corresponds to the non-central PI with proportion pair 4% and 94%. Such AW decline indicates that selecting proper proportion pair could effectively alleviate the interval width conservativeness. Nevertheless, the CCELM based proportion-free PIs achieve nearly 8% higher sharpness than the shortest PIs constructed by the MLLP model in Fig. 7. This sharpness enhancement arises from the CCELM's capability of generating both central

TABLE II
MODEL TRAINING TIME OF DIFFERENT FORECASTING METHODS

Method	SBL	BELM	BNN	HIA	MLLP	GKDE	FADiE	CCELM
Time (s)	0.06	1.37	2284.46	100.28	1.50	37.90	200.68	23.39

and non-central PIs with conditionally varying proportion pairs. Namely, the CCELM model could adaptively assign proper proportion pair to the PI at each time slot for the purpose of minimizing the overall interval width. To sum up, the CCELM based proportion-free PIs demonstrate remarkable potentialities for width shortening while ensuring high reliability.

C. Analysis of Computational Efficiency

Computational efficiency for model construction is critical to the development practicality of the probabilistic forecasting methods. Without loss of generality, the model training time for PIs with 95% NCP and 30-minute look-ahead time in spring is reported in Table II. All the applied forecasting models are implemented in MATLAB R2016b environment based on the PC with Intel Core i7-7700 CPU @ 3.60 GHz and 16 GB RAM. Since the persistence method is free of model training, its computation time is not investigated here.

In general, it takes more time to train nonparametric forecasting models in comparison with the parametric SBL and BELM models that fulfill model establishment in less than 1.5 seconds. Although these parametric models are sufficiently efficient, they inherently rely on distributional assumption to derive prediction intervals. As the neural network based parametric forecasting approaches, the model establishment of the BELM is 1666 times faster than the BNN. Actually, the computation relief of BELM approach is owing to the advanced scheme for data noise estimation and the utilization of ELM. As a result, training the BELM model is free of iterative gradient derivation and error backpropagation. Some nonparametric models such as the HIA and FADiE are fairly time-consuming, which require the training time up to several minutes. The HIA model is trained via particle swarm optimization, and the FADiE resorts to numerical approaches to solve linear partial differential equations in the density estimation stage. The GKDE method requires reasonable model training time owing to the analytical formulation of the predictive PDF. However, its forecasting performance is undesirable. As a nonparametric benchmark, the MLLP method achieves outstanding computational efficiency, but with inferior sharpness to the CCELM. The proposed CCELM method requires training time less than 25 seconds, which is about 76% and 88% more efficient than the HIA and FADiE methods respectively. In summary, as a nonparametric neuron network based model, the CCELM method possesses excellent computational efficiency and remarkably high potential for online applications.

D. Investigation of Training Algorithms

In the study, three state-of-the-art algorithms for solving chance constrained stochastic programming problem are utilized to validate the effectiveness of the proposed DCBS algorithm.

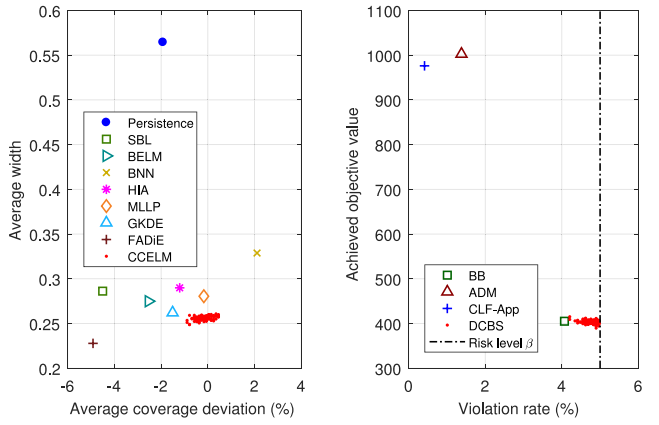
TABLE III
PERFORMANCE OF DIFFERENT ALGORITHMS FOR TRAINING CCELM MODEL

Algorithm	Achieved Objective	Violation Rate	Time (s)
BB	405.1131	4.07%	1800.46
ADM	1002.5070	1.38%	381.46
CLF-App	976.1173	0.41%	0.63
DCBS	403.7048	4.90%	26.24

Without loss of generality, PIs with 95% NCP and 1-hour look-ahead time in summer are studied for the CCELM model construction. The branch-and-bound (BB) algorithm implemented by the state-of-the-art commercial solver CPLEX is used to solve the mixed integer linear programming (MILP) reformulation of SAA problem (15) [36]. Such reformulation can also be tackled by the alternating direction method (ADM), which is able to find a suboptimal solution via sequentially solving convex quadratic programming and 0-1 linear knapsack problems with explicit solutions [45]. The indicator summation involved in the constraint (15c) can be replaced by the coherent loss function based inner convex approximation (CLF-App), which transforms the SAA problem into LP and is proved to be the tightest approximation among all the cumulative convex upper surrogates [46]. The convex quadratic programming problem in the ADM and LP problem in the CLF-App are tackled by off-the-shelf solvers MOSEK and CPLEX respectively.

Three criteria including the achieved objective value, violation rate of PI coverage for training dataset, and computational time are compared and presented in Table III. For the SAA problem (15) with risk level β being 5%, the achieved objective value should be as small as possible on condition that the violation rate does not exceed 5%. Actually, the BB algorithm achieves a suboptimal solution with the objective value 405.11 when the 30-minute time limit is reached. The objective values achieved by both the ADM and CLF-App are twice larger than that by the DCBS. Moreover, both the ADM and CLF-App have overly conservative violation rates less than 1.5%. The proposed DCBS algorithm spends only about 26 seconds achieving the smallest objective value 403.70, and simultaneously preserves the violation rate of PI coverage less than the allowed risk level. In summary, the proposed DCBS algorithm merits prominent superiority in terms of optimality attainment capability and computational time for solving the CCELM model.

In order to explore the influence of DCA's local optimality on the performance of the proposed CCELM model and DCBS algorithm, different initial solutions are fed into the DCA routine to examine the forecasting and optimization performance. To this end, the objective function $\mathbf{1}^\top \boldsymbol{\gamma}$ of the initial solution determination problem (26) is replaced by $\boldsymbol{w}^\top \boldsymbol{\gamma}$, where $\boldsymbol{w}$ is the weight coefficients randomly set as per the multivariate uniform distribution over $[0, 1]^{|\mathcal{T}|}$. Then the DCBS algorithm is executed repeatedly to obtain different converged solutions to the SAA problem (15). The forecasting and optimization performance metrics of the CCELM model and DCBS algorithm in 100 repetitions are recorded in Fig. 8(a) and 8(b) respectively. It can



(a) Forecasting performance (b) Optimization performance

Fig. 8. Performance comparisons of the CCELM model trained by DCBS algorithm given different initial solutions: (a) forecasting performance, (b) optimization performance.

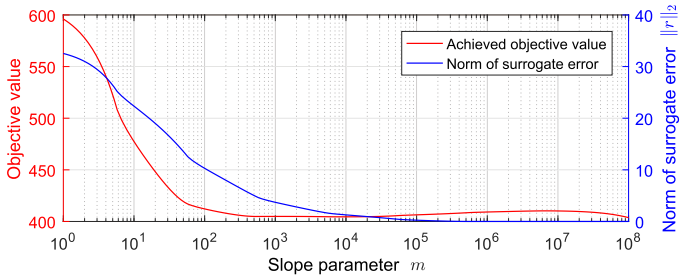


Fig. 9. Sensitivity analysis on the slope parameter m of the surrogate function.

be observed from Fig. 8(a) that the CCELM model maintains small ACDs varying from -0.89% to 0.45% and short AWs varying from 0.25 to 0.26, indicating excellent reliability and sharpness consistency. In particular, the red scatter points of the CCELM model are almost distributed in the bottom of Fig. 8(a) with ACDs close to zero. Therefore, the superior sharpness performance of the CCELM based PIs is on the premise of well reliability. As for the optimization performance metrics demonstrated in Fig. 8(b), the proposed DCBS algorithm achieves fairly competitive objective values. Note that the traditional BB algorithm fails to converge within the 30-minute time limit, which impedes practical applications. In Fig. 8(b), all the red scatters of the DCBS algorithm lie on the left of the risk level line, implying complete feasibility of attained solutions. To conclude, even though the DCA might produce local optimums, it would have negligible influence on the superior forecasting performance of the CCELM model and optimization performance of the DCBS algorithm.

In the DCBS algorithm, the parameter m controls the degree of resemblance between the surrogate and indicator function. Sensitivity analysis on the slope parameter is conducted. Under different slope parameter m, the objective value achieved by the DCBS algorithm and the ℓ_2-norm of the surrogate error $\|r\|_2$ with respect to all training samples are shown in Fig. 9. It can be observed that the norm of surrogate error and the achieved objective value drop slightly when m exceeds 10^3, which demonstrates the consistent performance within a wide range of slope parameter values. In general, the proposed DCBS algorithm behaves stably and the elaborated surrogate function achieves satisfactory accuracy in practical optimization process.

VI. Conclusion

Probabilistic wind power forecasting is crucial to power system decision-making under uncertainties. Traditional PIs rely on the predetermination of lower and upper bounds proportions, suffering from conservative sharpness. This paper develops an innovative chance constrained extreme learning machine model for constructing the optimal nonparametric PIs of wind power, which minimizes the expected interval width on the premise of well reliability. Benefited from the proportion-free model formulation and high nonlinear regression capability of ELM, the CCELM model has significant adaptivity and flexibility. The SAA counterpart of stochastic CCELM model is derived and proved to be equivalent to searching for the qualified parameter within polyhedral region pertaining to a parametric zero-one loss optimization problem. Difference of convex functions is designated to resemble the zero-one loss and enhance computational tractability. In order to train the developed CCELM model, a novel DC optimization based bisection search algorithm is proposed and efficiently accomplishes machine learning by solving a sequence of linear programming problems. Numerical studies based on actual wind farm with various confidence levels, look-ahead steps, seasonal patterns, and proportion pairs validate the significant effectiveness and efficiency of the CCELM model and DCBS algorithm. In general, the proposed methodology is promising to trustworthily support online model updating and assist various decision-making activities in power systems.

Appendix
Proofs of Propositions

A. Proof of Proposition 1

Proof: The proof relies on the following two lemmas.

Lemma 1 (Semicontinuity of Optimal Value Function [47]): The optimal value function $\rho(v)$ of a PO is lower semicontinous at v_0 if the corresponding feasible region $\Omega_{\text{PO}}(v_0)$ is compact and the objective function is lower semicontinuous on $\Omega_{\text{PO}}(v_0) \times \{v_0\}$.

Lemma 2 (Condition for Closed Sublevel Set [48]): The sublevel set of a function is closed if and only if the function is lower semicontinuous.

Since the indicator $\mathbb{I}(\gamma_t \leq 0)$ of the closed set $(-\infty, 0]$ is upper semicontinuous, the objective function (16a) formulated by its negative summation is lower semicontinuous. Provided that the feasible region $\Omega_{\text{PO}}(v)$ is bounded and closed, $\Omega_{\text{PO}}(v)$ is compact. According to *Lemma 1*, the optimal value function $\rho(v)$ is lower semicontinuous on its domain. Hence, it can be concluded that the $\beta|\mathcal{T}|$-sublevel set of $\rho(v)$, defined as

$$C_{\beta|\mathcal{T}|} := \{v \in \mathbb{R}_+ \mid \rho(v) \leq \beta|\mathcal{T}|\} \tag{27}$$

is closed by applying *Lemma 2*.

Recalling the feasibility of SAA problem (15), the $\beta|\mathcal{T}|$-sublevel set (27) of the optimal value function $\rho(v)$ is therefore

a nonempty closed interval and the left endpoint of interval is exactly the goal parameter value $v = \hat{v}$ such that $\rho(v) \leq \beta|\mathcal{T}|$. This establishes *Proposition 1*. ■

B. Proof of Proposition 2

Proof: For the SAA counterparts (15) of the stochastic CCELM model, let $(\gamma_t^\star, \boldsymbol{\omega}_\ell^\star, \boldsymbol{\omega}_u^\star)$ be the minimizer. For the PZO (16), let $\hat{v}$ be the minimum parameter v such that $\rho(v) \leq \beta|\mathcal{T}|$, and $(\hat{\gamma}_t, \hat{\boldsymbol{\omega}}_\ell, \hat{\boldsymbol{\omega}}_u)$ be the corresponding minimizer.

Note that the constraints (15b) and (15d) of the SAA counterpart (15) are identical to the constraints (16b) and (16d) of the PZO problem (16). Given the definition of parameter value $\hat{v}$, the inequality $\rho(\hat{v}) \leq \beta|\mathcal{T}|$ holds and constraint (15c) is therefore satisfied at $\gamma_t = \hat{\gamma}_t$. Trivially, the relationship $(\hat{\gamma}_t, \hat{\boldsymbol{\omega}}_\ell, \hat{\boldsymbol{\omega}}_u) \in \Omega_{\text{SAA}}$ can be deduced. Since a feasible solution only achieves objective value no less than the optimal value, it can be deduced that the inequality $\hat{v} \geq v^\star$ holds for the SAA problem (15). Then proving $\hat{v} = v^\star$ suffices to show the absurdity of $\hat{v} > v^\star$.

If the strict inequality $\hat{v} > v^\star$ holds, the relationship $(\gamma_t^\star, \boldsymbol{\omega}_\ell^\star, \boldsymbol{\omega}_u^\star) \in \Omega_{\text{PO}}(\hat{v})$ can be deduced due to the feasibility of constraint (16c) for the PZO problem (16). This gives rise to the inequalities below

$$\rho(v^\star) \leq \sum_{t\in\mathcal{T}} [1 - \mathbb{I}\left(\gamma_t^\star \leq 0\right)] \leq \beta|\mathcal{T}| \tag{28}$$

The first inequality in (28) is valid since the feasible solution $(\gamma_t^\star, \boldsymbol{\omega}_\ell^\star, \boldsymbol{\omega}_u^\star)$ achieves objective value $\sum_{t\in\mathcal{T}}[1 - \mathbb{I}(\gamma_t^\star \leq 0)]$ no less than the optimal value $\rho(v^\star)$ for the PZO (16). The second inequality follows due to the constraint (15c). The inequalities (28) mean that there exists a smaller parameter value $v^\star$ than $\hat{v}$ for the PZO problem (16) such that $\rho(v) \leq \beta|\mathcal{T}|$, which contradicts the definition of parameter value $\hat{v}$. Thus, only the case $\hat{v} = v^\star$ is valid for the previously deduced inequality $\hat{v} \geq v^\star$, which concludes the proof of *Proposition 2*. ■

C. Proof of Proposition 3

Proof: Given any fixed parameter $v \in \mathbb{R}_+$, let $(\underline{\gamma}_t, \underline{\boldsymbol{\omega}}_\ell, \underline{\boldsymbol{\omega}}_u)$ be the optimal solution to the relaxed PZO problem (17). Due to the inequality relaxation (17b)–(17c), it holds that

$$\underline{\rho}(v) \leq \rho(v) \tag{29}$$

One can easily utilize the relaxed solution $(\underline{\boldsymbol{\omega}}_\ell, \underline{\boldsymbol{\omega}}_u)$ to recover a feasible solution $(\gamma_t', \underline{\boldsymbol{\omega}}_\ell, \underline{\boldsymbol{\omega}}_u)$ to the original PZO problem (16) as per

$$\gamma_t' := \max\{f(\boldsymbol{x}_t, \underline{\boldsymbol{\omega}}_\ell) - y_t, y_t - f(\boldsymbol{x}_t, \underline{\boldsymbol{\omega}}_u)\} \tag{30}$$

Since a feasible solution always achieves the objective value greater than or equal to the optimal value, it is valid that

$$\sum_{t\in\mathcal{T}} [1 - \mathbb{I}\left(\gamma_t' \leq 0\right)] \geq \rho(v) \tag{31}$$

Note that the inequality $\gamma_t' \leq \underline{\gamma}_t$ is valid due to the constraints (17b) and (17c). If $\gamma_t' \leq 0$, then the inequality $\gamma_t' \leq \underline{\gamma}_t \leq 0$ must be satisfied in order to minimize the zero-one loss (17a). If $\gamma_t' > 0$, then the inequality $\underline{\gamma}_t \geq \gamma_t' > 0$ can be deduced. Consequently, $\underline{\gamma}_t$ is with the same sign as γ_t', namely,

$$\underline{\rho}(v) = \sum_{t\in\mathcal{T}} [1 - \mathbb{I}(\underline{\gamma}_t \leq 0)] = \sum_{t\in\mathcal{T}} [1 - \mathbb{I}\left(\gamma_t' \leq 0\right)] \geq \rho(v) \tag{32}$$

Combining the inequality (29) and inequality (32) implies $\underline{\rho}(v) = \rho(v)$. Hence, the exactness of such inequality relaxation follows. ■

REFERENCES

[1] K. Ohlenforst *et al.*, "Global wind report 2018," Global Wind Energy Council, Brussels, Belgium, Tech. Rep., Apr. 2019.

[2] C. Wan, J. Lin, W. Guo, and Y. Song, "Maximum uncertainty boundary of volatile distributed generation in active distribution network," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 2930–2942, Jul. 2018.

[3] N. Zhang, C. Kang, Q. Xia, and J. Liang, "Modeling conditional forecast error for wind power in generation scheduling," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1316–1324, May 2014.

[4] Y. Jiang, C. Wan, J. Wang, Y. Song, and Z. Y. Dong, "Stochastic receding horizon control of active distribution networks with distributed renewables," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1325–1341, Mar. 2019.

[5] J. Yan, F. Li, Y. Liu, and C. Gu, "Novel cost model for balancing wind power forecasting uncertainty," *IEEE Trans. Energy Convers.*, vol. 32, no. 1, pp. 318–329, Mar. 2017.

[6] R. Doherty and M. O'Malley, "A new approach to quantify reserve demand in systems with significant installed wind capacity," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 587–595, May 2005.

[7] H. Bludszuweit, J. A. Domínguez-Navarro, and A. Llombart, "Statistical analysis of wind power forecast error," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 983–991, Aug. 2008.

[8] K. Bruninx and E. Delarue, "A statistical description of the error on wind power forecasts for probabilistic reserve sizing," *IEEE Trans. Sustain. Energy*, vol. 5, no. 3, pp. 995–1002, Jul. 2014.

[9] C. Wan, Y. Song, Z. Xu, G. Yang, and A. H. Nielsen, "Probabilistic wind power forecasting with hybrid artificial neural networks," *Elect. Power Compon. Syst.*, vol. 44, no. 15, pp. 1656–1668, Sep. 2016.

[10] T. Heskes, "Practical confidence and prediction intervals," in *Proc. 9th Int. Conf. Neural Inf. Process. Syst.*, Denver, Colorado, USA, Dec. 1996, pp. 176–182.

[11] A. Khosravi, S. Nahavandi, and D. Creighton, "Prediction intervals for short-term wind farm power generation forecasts," *IEEE Trans. Sustain. Energy*, vol. 4, no. 3, pp. 602–610, Jul. 2013.

[12] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, "Probabilistic forecasting of wind power generation using extreme learning machine," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1033–1044, May 2014.

[13] P. Pinson and G. Kariniotakis, "Conditional prediction intervals of wind power generation," *IEEE Trans. Power Syst.*, vol. 25, no. 4, pp. 1845–1856, Nov. 2010.

[14] G. Sideratos and N. D. Hatziargyriou, "Probabilistic wind power forecasting using radial basis function neural networks," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 1788–1796, Nov. 2012.

[15] A. U. Haque, M. H. Nehrir, and P. Mandal, "A hybrid intelligent model for deterministic and quantile regression approach for probabilistic wind power forecasting," *IEEE Trans. Power Syst.*, vol. 29, no. 4, pp. 1663–1672, Jul. 2014.

[16] C. Wan, Z. Xu, and P. Pinson, "Direct interval forecasting of wind power," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4877–4878, Nov. 2013.

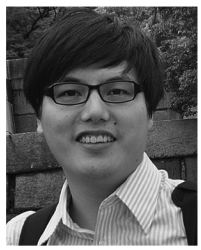
[17] J. W. Taylor, P. E. McSharry, and R. Buizza, "Wind power density forecasting using ensemble predictions and time series models," *IEEE Trans. Energy Convers.*, vol. 24, no. 3, pp. 775–782, Sep. 2009.

[18] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, "Optimal prediction intervals of wind power generation," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1166–1174, May 2014.

[19] R. J. Bessa, V. Miranda, A. Botterud, J. Wang, and E. M. Constantinescu, "Time adaptive conditional kernel density estimation for wind power forecasting," *IEEE Trans. Sustain. Energy*, vol. 3, no. 4, pp. 660–669, Oct. 2012.

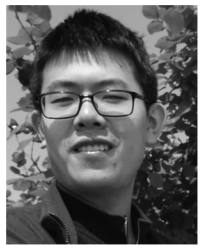
[20] C. Wan, J. Lin, J. Wang, Y. Song, and Z. Y. Dong, "Direct quantile regression for nonparametric probabilistic forecasting of wind power generation," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2767–2778, Jul. 2017.

[21] Y. Lin, M. Yang, C. Wan, J. Wang, and Y. Song, "A multi-model combination approach for probabilistic wind power forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 226–237, Jan. 2019.
[22] B. Khorramdel, C. Y. Chung, N. Safari, and G. C. D. Price, "A fuzzy adaptive probabilistic wind power prediction framework using diffusion kernel density estimators," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7109–7121, Nov. 2018.
[23] W. Xie, P. Zhang, R. Chen, and Z. Zhou, "A nonparametric Bayesian framework for short-term wind power probabilistic forecast," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 371–379, Jan. 2019.
[24] J. Toubeau, J. Bottieau, F. Valle, and Z. D. Grve, "Deep learning-based multivariate probabilistic forecasting for short-term scheduling in power markets," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1203–1215, Mar. 2019.
[25] H. Hoeltgebaum, C. Fernandes, and A. Street, "Generating joint scenarios for renewable generation: The case for non-Gaussian models with time-varying parameters," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7011–7019, Nov. 2018.
[26] C. Wan, Z. Xu, Y. Wang, Z. Y. Dong, and K. P. Wong, "A hybrid approach for probabilistic forecasting of electricity price," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 463–470, Jan. 2014.
[27] F. Golestaneh, P. Pinson, R. Azizipanah-Abarghooee, and H. B. Gooi, "Ellipsoidal prediction regions for multivariate uncertainty characterization," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 4519–4530, Jul. 2018.
[28] Z. Cao, C. Wan, Z. Zhang, F. Li, and Y. Song, "Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 1881–1897, May. 2020.
[29] C. Wan, J. Lin, Y. Song, Z. Xu, and G. Yang, "Probabilistic forecasting of photovoltaic generation: An efficient statistical approach," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2471–2472, May 2017.
[30] C. Wan, J. Wang, J. Lin, Y. Song, and Z. Y. Dong, "Nonparametric prediction intervals of wind power via linear programming," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 1074–1076, Jan. 2018.
[31] C. Zhao, C. Wan, and Y. Song, "An adaptive bilevel programming model for nonparametric prediction intervals of wind power generation," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 424–439, Jan. 2020.
[32] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Statistical Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007.
[33] R. Askanazi, F. X. Diebold, F. Schorfheide, and M. Shin, "On the comparison of interval forecasts," *J. Time Series Anal.*, vol. 39, no. 6, pp. 953–965, Nov. 2018.
[34] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, Dec. 2006.
[35] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, "General conditions for predictivity in learning theory," *Nature*, vol. 428, no. 6981, pp. 419–422, Mar. 2004.
[36] J. Luedtke, "A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support," *Math. Program.*, vol. 146, no. 1, pp. 219–244, Aug. 2014.
[37] P. Bonami, A. Lodi, A. Tramontani, and S. Wiese, "On mathematical programming with indicator constraints," *Math. Program.*, vol. 151, no. 1, pp. 191–223, Jun. 2015.
[38] Y. Wu and Y. Liu, "Robust truncated hinge loss support vector machines," *J. Amer. Statistical Assoc.*, vol. 102, no. 479, pp. 974–983, 2007.
[39] F. Pérez-Cruz, A. Navia-Vázquez, A. R. Figueiras-Vidal, and A. Artes-Rodriguez, "Empirical risk minimization for support vector classifiers," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 296–303, Mar. 2003.
[40] L. T. H. An and P. D. Tao, "The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems," *Ann. Oper. Res.*, vol. 133, no. 1, pp. 23–46, Jan. 2005.
[41] Actual generation output per generation unit. Balancing Mechanism Reporting Service (BMRS) of ELEXON Limited, 2017. [Online]. Available: https://www.bmreports.com/bmrs/?q=actgenration/actualgeneration. Accessed on: Dec. 13, 2018.
[42] P. Pinson, H. A. Nielsen, J. K. Mller, H. Madsen, and G. N. Kariniotakis, "Non-parametric probabilistic forecasts of wind power: Required properties and evaluation," *Wind Energy*, vol. 10, no. 6, pp. 497–516, 2007.
[43] C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, "Wind power forecasting: State-of-the-art 2009." Argonne National Laboratory, Argonne, IL, USA, Tech. Rep. ANL/DIS-10-1 TRN US200924351, 2009.
[44] C. Wan, M. Niu, Y. Song, and Z. Xu, "Pareto optimal prediction intervals of electricity price," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 817–819, Jan. 2017.
[45] X. Bai, J. Sun, X. Sun, and X. Zheng, "An alternating direction method for chance-constrained optimization problems with discrete distributions," School of Management, Fudan University, Shanghai, China, Tech. Rep., 2012.
[46] W. Yang, M. Sim, and H. Xu, "Goal scoring, coherent loss and applications to machine learning," *Math. Program.*, 2019, to be published, doi: 10.1007/s10107-019-01387-y.
[47] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Non-Linear Parametric Optimization*. Basel, Switzerland: Birkhäuser, 1983.
[48] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, 4th ed. Dordrecht, Netherlands: Springer, 2012.



Can Wan (Member, IEEE) received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2008, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2015.

He serves as a Research Professor with the College of Electrical Engineering, Zhejiang University, Hangzhou, China, under the University Hundred Talents Program. He was a Postdoctoral Fellow with the Department of Electrical Engineering, Tsinghua University, Beijing, China, and held research positions at the Technical University of Denmark, The Hong Kong Polytechnic University, and City University of Hong Kong. He was a Visiting Scholar with the Center for Electric Power and Energy, Technical University of Denmark, and Argonne National Laboratory, Lemont, IL, USA. His research interests include forecasting, renewable energy, active distribution network, integrated energy systems, and machine learning. He is an Associate Editor for the IEEE SYSTEMS JOURNAL.



Changfei Zhao received the B.Eng. degree in electrical engineering from the School of Automation, Nanjing University of Science and Technology, Nanjing, China, in 2017. He is currently working toward the Ph.D. degree with the College of Electrical Engineering, Zhejiang University, Hangzhou, China.

He serves as a Research Assistant with the Smart Grid Operation and Optimization Laboratory, College of Electrical Engineering, Zhejiang University. His research interests lie in power system uncertainty analysis, decision-making under uncertainties, mathematical optimization theory, and statistical learning.

Yonghua Song (Fellow, IEEE) received the B.Eng. and Ph.D. from the Chengdu University of Science and Technology (now Sichuan University) and China Electric Power Research Institute in 1984 and 1989 respectively.

He is currently the Rector of University of Macau and the Director of State Key Laboratory of Internet of Things for Smart City, and also an Adjunct Professor with the Department of Electrical Engineering, Tsinghua University, and the College of Electrical Engineering, Zhejiang University. He was awarded D.Sc., Honorary D.Eng. and Honorary D.Sc. by Brunel University, University of Bath and University of Edinburgh, respectively. He is a Fellow of the Royal Academy of Engineering and a Foreign Member of Academia Europaea.