IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 May 2024, accepted 3 June 2024, date of publication 7 June 2024, date of current version 17 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3411400

RESEARCH ARTICLE

Robust Kernel Density Estimation Based Data-Driven Optimal Scheduling for Power Systems Considering Data Errors and Uncertainties of Renewable Energy

WENTING HOU[1], LONGXIAN YI[2], HUAIBIN MIAO[1], AND YINING MA[3]

[1]School of Mechanical and Electrical Engineering, Zhoukou Normal University, Zhoukou 466001, China
[2]Nanning Branch of Guangxi Radio and Television Technology Center, Nanning 530016, China
[3]School of Mechanical Engineering, University of Leeds, LS2 9JT Leeds, U.K.

Corresponding author: Wenting Hou (houwenting99@163.com)

This work was supported in part by the Key Scientific Research Project of Colleges and Universities in Henan Province under Grant 23B470003, in part by the Key Research and Development and Promotion Special Project of Henan Province under Grant 232102311228 and Grant 212102210246, and in part by Zhoukou Normal University under Grant ZKNUC2019009.

ABSTRACT This paper proposes a data-driven robust scheduling method for power systems incorporating variable energy. Robust kernel density estimation (RKDE) is combined with distributionally robust optimization (DRO) to address the uncertainties of renewable energy and possible outliers during data collection and transmission. RKDE is employed to infer the potential probability distribution. In this process, the outliers will be assigned a very small weight so that they hardly affect the probability density curve. Subsequently, the distribution derived from RKDE serves as the center of a distributional ambiguity set, with distances between distributions measured using the Wasserstein metric. Since RKDE converges to the true distribution quickly with the expansion of sample data, the proposed approach is less conservative than the empirical distribution-based DRO (EDRO). Moreover, compared with general KDE, RKDE has a unique advantage in suppressing the influence of outliers and improving the accuracy of distribution estimation. To demonstrate the superiority of the proposed approach, we present tests on Case-118 and Case-1888rte systems from MATPOWER 6.0. Numerical results indicate that the proposed approach exhibits lower conservatism and superior outlier suppression capability when compared to EDRO and KDE-based DRO (KDRO).

INDEX TERMS Data-driven, robust optimization, power system scheduling, renewable energy, uncertainty, data errors.

NOMENCLATURE

Sets

I, J	Set of thermal units and renewable generators.
I_b, J_b	Set of all thermal units and renewable generators at bus b.
B, L, T	Set of buses, transmission lines and time periods.

Parameters

a_i, b_i, c_i	Generation cost coefficients of unit i.
C_{mn}	Capacity of transmission line linking bus m and n.
CU_i/CD_i	Upward/downward reserve price of unit i.
K^b_{mn}	Load shift factor from bus b to line linking bus m and n.
PL^b_t	Load at bus b in time t.
P^w_{jt}	Predicted value of renewable source j in time t.
$\bar{P}_i/\underline{P}_i$	Maximum/minimum output of unit i.

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Bellan.

RU_i/RD_i	Ramp-up/ramp-down limit of unit i.
$\overline{RU}_i/\overline{RD}_i$	Start-up ramp-up/shut-down ramp-down limit of unit i.
SU_i/SD_i	Start-up/shut-down cost of unit i.
TU_i/TD_i	Minimum up/down time of unit i.
$\bar{r}_{it}/r_{it}$	Upward/downward reserve limit of unit i.

Variables

d_{it}	On/off status of units: '1'if unit i is on, '0' if unit i is off.
P_{it}	Base-point output of unit i in time t.
u_{it}	Start-up status of units: '1'if unit i is started up in time t, '0' otherwise.
v_{it}	Shut-down status of unit s: '1'if unit i is shut down in time t, '0' otherwise.
α_{it}	Participation factor of unit i in time t.
$\tilde{\zeta}_{jt}$	Prediction error of renewable source j in time t.
ζ_t	Total prediction error of all renewable sources in time t.

I. INTRODUCTION

In recent years, volatile renewable energy is making up a rapidly growing share in power systems. The resulting uncertainties bring enormous risks to the stable operation of power systems [1]. In addition to the various uncertainties, the inevitable data errors in the process of renewable energy data acquisition and transmission also bring great difficulties to accurate power scheduling. Exploring approaches to address the above issues is a constant challenge for researchers.

For the optimal scheduling of power systems under uncertainties, most scholars focus on coping with uncertainty [2] but rarely consider the data errors. To deal with uncertainties, stochastic programming (SP) [3]is one of the popular methods that researchers have invested in. It usually assumes uncertainties following a typical distribution and solves the model by constructing massive scenarios [4] or chance constraints [5]. In contrast with SP, robust optimization (RO) does not need to know what probability distribution the uncertainties follow [6], [7], but constructs a set containing the worst scenario [8]. The objective is to minimize the cost under the worst realization [9]. Both of the above two methods can effectively describe the uncertainties, but SP is aggressive while RO is conservative [10], [11], which limits their effectiveness and economics when applied to actual scheduling problems.

To strike a balance between aggressive and conservative performance, stochastic scenarios are integrated with robust optimization in [12]. Subsequently, researchers propose distributionally robust optimization (DRO), which focuses on the utilization of historical data information, such as the cumulative distribution [13], mean and variance [14], etc. In [15], DRO is integrated with interval optimization to solve the optimal power flow for integrated electricity and natural gas systems. In [16], the N-k security criterion and moment information of contingency are used to form a distributionally robust contingency-constrained framework for unit commitment. In [17], the Kullback-Leibler divergence is applied to establish a min-max-min distributionally robust model. In [18], the Wasserstein metric is used to construct a data-driven distributionally robust unit commitment model.

The mining and utilization of historical data is very common in SP, RO and DRO. In addition to the above measures, nonparametric statistical methods are often introduced into the field of uncertainty characterization, e.g. kernel density estimation (KDE). KDE is a well-known nonparametric approach to estimate the potential distribution of stochastic load or wind power. It has been widely used in probabilistic load flow [19], optimal stochastic scheduling [20], short-term wind power prediction [21], load curve classification [22], etc. Beyond that, to address the optimal scheduling operation of power systems with high penetration of renewable, KDE is applied to interval optimization in [23]. In [24], KDE-based robust optimization is applied to the problem of electric vehicle charging station location, it is used to reduce the conservatism of the robust optimization. In [25], KDE is combined with clustering to form a data-driven stochastic robust optimization approach for sustainable utility systems. Since KDE is determined based on historical data, the error caused by selecting a particular probability distribution can be avoided. Moreover, KDE has the ability to converge to the true distribution, thereby reducing the conservatism of robust optimization.

The methods mentioned above rely on historical data, assuming it is completely accurate. However, during the process of collecting and transmitting massive amounts of data, there will inevitably be some errors (i.e., contaminated samples), which can impact the accuracy of these methods. Considering this situation, this paper proposes a robust kernel density estimation (RKDE) [26] based distributionally robust scheduling method for power systems under uncertainty. RKDE could estimate the true distribution as much as possible while suppressing the outliers. On the one hand, RKDE inherits the advantages of KDE, that is, it can converge to the real probability density quickly as the data size increases. On the other hand, different from KDE, RKDE assigns different weight values to the kernel functions corresponding to the sample data. In this process, the outliers will be assigned a smaller weight, so as to ensure the accuracy of the probability density estimation. In view of this, the combination of RKDE and the Wasserstein metric is helpful in constructing a compact distributional ambiguity set and achieving a reasonable description of renewable energy uncertainty. Crucially, it is beneficial to reduce the conservatism of the model and improve the immunity to outliers.

The main contributions of this research work are aggregated as follows:

(1) A RKDE-based data-driven distributionally robust scheduling method for power systems is developed in this paper. RKDE is adopted to mine the probabilistic information of renewable energy. In contrast with general KDE, it is able

to weaken the influence of outliers in the historical data and deduce the reliable probability density.

(2) RKDE is integrated with Wasserstein distance. The distances of distributions are measured by the Wasserstein metric, and the distribution estimated by RKDE is taken as the center of the distributional ambiguity set, which is constructed based on the support space deduced by RKDE. The proposed approach ensures the convergence of the ambiguity set towards the real distribution with less conservatism than empirical distribution-based DRO.

(3) The proposed method, namely RKDE-based DRO (RDRO), is compared with empirical distribution-based DRO (EDRO) and KDE-based DRO (KDRO) using test systems. Simulation results reveal that the presented method is more reliable and less conservative. Besides, it has a good ability to suppress the outliers.

The rest of the paper is organized as follows. Section II depicts the proposed approach, the construction of support space and the distributional ambiguity set. Section III describes the construction and reformulation of the proposed scheduling model. Section IV presents numerical results on the Case-118 and Case-1888rte systems from MATPOWER 6.0. Finally, the conclusions are drawn in Section V.

II. DATA-DRIVEN SCHEDULING METHOD WITH OUTLIER SUPPRESSION CAPABILITY

A. FRAMEWORK OF THE PROPOSED DATA-DRIVEN APPROACH

In empirical distribution-based DRO (EDRO) [27], the support of uncertainties is obtained through inference. The stochastic variables are first converted into standard forms via sample mean and variance. Then the minimum interval of standard forms satisfying a given confidence value is calculated, and the support space is deduced by this. Nevertheless, this approach will lead to a loose estimate of the limits of uncertain variables, resulting in an excessively conservative support space. Conversely, RKDE can converge to the real probability density as the data size increases. In this paper, it is used to extract the distribution from the historical data of renewable energy. A more compact support could be constructed by the confidence interval of the derived distribution.

The flow chart of the proposed data-driven approach with outlier suppression capability is shown in Fig. 1. Cumulative density is extracted from the contaminated sample to serve as the central distribution of the distributional ambiguity set. In this process, the outliers contribute very little because they are assigned a small weight by RKDE. Since the derived cumulative density can almost be regarded as the real distribution, its confidence interval can be used as the support space of uncertainties to reduce conservatism. Using the Wasserstein distance as a measure between distributions, an ambiguity set of distributions with low conservatism and immunity data anomalies can be constructed.

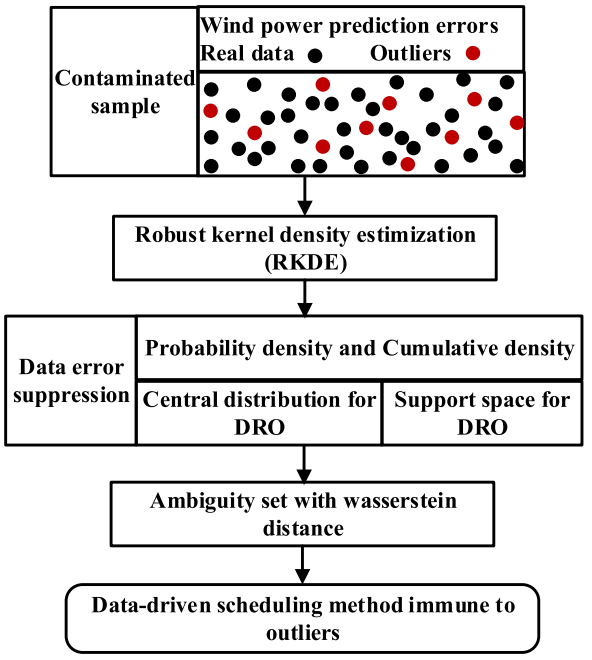

FIGURE 1. Flow chart of the proposed data-driven approach.

B. SUPPORT SPACE OF VOLATILE RENEWABLE ENERGY

The probability density of renewable energy generation is

$$f_{RKDE}(\xi) = \sum_{i=1}^{N} w_i K_h(\xi, \xi_i),\ \ \xi_i \in R^d = \sum_{i=1}^{N} w_i \left(2\pi h^2\right)^{-d/2} \exp\left(-\frac{\|\xi - \xi_i\|^2}{2h^2}\right) \quad (1)$$

where $f_{RKDE}(\cdot)$ is the probability density function deduced by RKDE, ξ represents uncertainties, as the prediction errors of renewable energy, K_h is a Gaussian kernel with a bandwidth h, and w_i is a weight factor of Gaussian kernels corresponding to the samples.

For RKDE, the bandwidth and weights of Gaussian kernels are very important in the derivation of probability density. This paper adopts the biased cross-validation approach (BCVA) to calculate the optimal value of the bandwidth. The details of BCVA are shown below.

$$BCVA(h) = \sqrt{\frac{1}{(4\pi n^2 h^2)}} + \frac{h^4}{4n^2} \sum_{i(i\neq j)}^{N} \sum_{j}^{N} [K^{(4)}_{\sqrt{2}h}(\xi_i - \xi_j)] \quad (2)$$

where $K_b^{(d)}(\cdot)$ represents a d-dimensional Gaussian kernel with a bandwidth b as

$$K^{(4)}_{\sqrt{2}\,h}(\xi) = \left[\frac{\xi^4}{(\sqrt{2}h)^9} - \frac{6\xi^2}{(\sqrt{2}\,h)^2} + \frac{3}{(\sqrt{2}h)^5}\right] \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{4h^2}} \quad (3)$$

The optimal value of bandwidth $\tilde{h}$ is determined by minimizing $BCVA(h)$.

$$\tilde{h} = \arg\min_{h} BCVA(h) \tag{4}$$

The weights $\hat{w}_i$ of Gaussian kernels are calculated by the kernelized iteratively re-weighted least squares (KIRWLS) approach. For details, please see the [26]. After that, we get the probability density function

$$f_{RKDE}(\xi) = \sum_{i=1}^{N} \hat{w}_i K_h(\xi, \xi_i) \tag{5}$$

Then the cumulative density $F_{RKDE}(\xi)$, alpha quantile function $F^{-1}_{RKDE}(\alpha)$ and the support space of renewable energy $\mathbb{S}$ can be expressed as

$$F_{RKDE}(\xi) = \int_{-\infty}^{+\infty} f_{RKDE}(\xi) d\xi \tag{6}$$

$$F^{-1}_{RKDE}(\alpha) = \min\{\xi \in R | F_{RKDE}(\xi) \geq \alpha\} \tag{7}$$

$$\mathbb{S} = \left\{ \boldsymbol{\xi} \mid \begin{array}{l} \underline{\xi} \leq \xi_i \leq \bar{\xi}, \underline{\xi} = F^{-1}_{RKDE}(\frac{1-\alpha}{2}) \\ \bar{\xi} = F^{-1}_{RKDE}(\frac{1+\alpha}{2}),\ \forall i \in [1, N] \end{array} \right\} \tag{8}$$

where $\bar{\xi}$, $\underline{\xi}$ represent the upper and lower limits of ξ_i, α indicates the confidence level.

C. AMBIGUITY SET OF DISTRIBUTIONS

The ambiguity set consists of distributions, the distances between which are measured by the Wasserstein metric. For a sample set containing N prediction errors of renewable energy $\{\xi_1, \xi_2, \cdots, \xi_N\}$, the Wasserstein distance between the RKDE-based distribution F_{RKDE} and the true distribution F_{true} can be defined as

$$\mathbb{W}(F_{RKDE}, F_{true}) = \inf_{\Pi} \left\{ \int \mathbb{S} d(\xi, \tilde{\xi}) \Pi(d\xi, d\tilde{\xi}), \right.$$
$$\left. \xi \sim F_{RKDE}, \tilde{\xi} \sim F_{true} \right\} \tag{9}$$

where Π represents the joint distribution of ξ and $\tilde{\xi}$.

A set of distributions over a distance d_w and centered at the RKDE-based distribution constitute the ambiguity set $\mathbb{F}$.

$$\mathbb{F} = \{F \in \Re(\mathbb{S}) | \mathbb{W}(F_{RKDE}, F) \leq d_w\} \tag{10}$$

Given a confidence level ℓ, the value of distance d_w could be calculated by the following steps [27].

$$F[\mathbb{W}(F_{RKDE}, F) \leq d_w]$$
$$\geq 1 - \exp(-N d_w^2 / \varepsilon^2) \tag{11}$$

$$\ell = 1 - \exp(-N d_w^2 / \varepsilon^2) \tag{12}$$

$$d_w = \varepsilon(\ln((1-\ell)^{-1})/N)^{0.5} \tag{13}$$

$$\varepsilon \approx 2 \min_{\lambda > 0, \lambda \in R} \left(\frac{1}{2\lambda} \left(1 + \ln(N^{-1} \sum_{i=1}^{N} e^{\lambda \| \xi_i - \mu \|^2}) \right) \right)^{0.5} \tag{14}$$

where μ is the sample mean and λ is a positive real number.

III. SCHEDULING MODEL CONSTRUCTION AND REFORMULATION

A min-max model for power scheduling under uncertainty is established, which aims to minimize the operation costs under the worst distribution.

$$\min SU_i u_{it} + SD_i v_{it} + \max_{F \in F_s} \mathrm{E}_F(CU_i \bar{r}_{it} + CD_i r_{it} + f(P, \zeta)) \tag{15}$$

$$s.t.\ f(P, \zeta) = a_i(P_{it} + \alpha_{it}\zeta_t)^2 + b_i(P_{it} + \alpha_{it}\zeta_t) + c_i) \tag{16}$$

$$TU_i(d_{it} - d_{i(t-1)}) \leq \sum_{k=0}^{TU_i - 1} d_{i(t+k)},\ \forall i \in I,\ \forall t \in T \tag{17}$$

$$TD_i(d_{it} - d_{i(t-1)}) \leq \sum_{k=0}^{TD_i - 1} (1 - d_{i(t+k)}),\ \forall i \in I,\ \forall t \in T \tag{18}$$

$$d_{it} - d_{i(t-1)} - u_{it} \leq 0,\ \forall i \in I,\ \forall t \in T \tag{19}$$

$$-d_{it} + d_{i(t-1)} - v_{it} \leq 0,\ \forall i \in I,\ \forall t \in T \tag{20}$$

$$d_{it}, u_{it}, v_{it} \in \{0, 1\},\ \forall i \in I,\ \forall t \in T \tag{21}$$

$$\sum_{i \in I} (P_{it} + \alpha_{it}\zeta_t) + \sum_{j \in J} (P^w_{jt} + \tilde{\zeta}_{jt}) = \sum_{b \in B} PL^b_t,\ \forall t \in T \tag{22}$$

$$d_{it} P_i + r_{it} \leq P_{it} \leq d_{it} \bar{P}_i - \bar{r}_{it},\ \forall i \in I,\ \forall t \in T \tag{23}$$

$$(P_{it} + \bar{r}_{it}) - (P_{i(t-1)} - r_{i(t-1)}) \leq (2 - d_{i(t-1)} - d_{it})$$
$$\overline{RU}_i + (1 + d_{i(t-1)} - d_{it}) RU_i,\ \forall i \in I,\ \forall t \in T \tag{24}$$

$$(P_{i(t-1)} + \bar{r}_{i(t-1)}) - (P_{it} - r_{it}) \leq (2 - d_{i(t-1)} - d_{it})$$
$$\overline{RD}_i + (1 - d_{i(t-1)} + d_{it}) RD_i,\ \forall i \in I,\ \forall t \in T \tag{25}$$

$$\sum_{i \in I} \alpha_{it} = 1,\ \ 0 \leq \alpha_{it} \leq s_{it},\ \forall i \in I,\ \forall t \in T \tag{26}$$

$$-r_{it} \leq \alpha_{it}\zeta_t \leq \bar{r}_{it}, \bar{r}_{it},\ r_{it} \geq 0,\ \forall i \in I,\ \forall t \in T \tag{27}$$

$$\zeta_t = \sum_{j \in J} \tilde{\zeta}_{jt},\ \forall t \in T \tag{28}$$

$$-C_{mn} \leq \sum_{b \in B} K^b_{mn} \left(\sum_{i \in I_b} (P_{it} + \alpha_{it}\zeta_t) + \sum_{j \in J_b} (P^w_{jt} + \tilde{\zeta}_{jt}) - PL^b_t \right) \leq C_{mn},\ \forall i \in I,\ \forall t \in T, (i, j) \in L \tag{29}$$

Constraint (15) aims to minimize the start-up and shut-down costs, the expected reserve and generation costs of units. Constraints (16) represent the generation cost of thermal units. Formulas (17) and (18) indicate the minimum up-time and down-time of thermal units. Equations (19) and (20) indicate the mathematical relationship of three binary variables. Constraints (22) represent the balance between load and power supply. Constraints (23) enforce the output limits of thermal units. Constraints (24) describe the ramp-up limits for unit start-up and continuous operation. Constraints (25) describe the ramp-down limits for unit shut-down and continuous operation. Constraints (26) indicate that the sum of participation factors of all online thermal units is equal to one. Constraints (27) indicate that

the renewable energy power fluctuation borne by thermal units cannot exceed the reserve capacity. Constraint (28) is the total prediction error of all renewable sources at time t. Constraints (29) are the line limits of transmission power.

To facilitate the calculation of the proposed model, the expected reserve and generation costs under the worst distribution are reformulated into the following form [13].

$$\begin{cases} f(\zeta_t) = a'\zeta_t^2 + b'\zeta_t + c' \\ = CU_i\bar{r}_{it} + CD_i r_{it} + f(P,\zeta) \\ a' = \sum_{t\in T}\sum_{i\in I} a_i\alpha_{it}^2 \\ b' = \sum_{t\in T}\sum_{i\in I} 2a_iP_{it}\alpha_{it} + b_i\alpha_{it} \\ c' = \sum_{t\in T}\sum_{i\in I} (a_iP_{it}^2 + b_iP_{it} + c_i) + CU_i\bar{r}_{it} + CD_i r_{it} \end{cases} \tag{30}$$

Then the equivalent expression in "min" form is obtained as

$$\begin{aligned} &\max_{F\in F_s} \mathrm{E}_F(CU_i\bar{r}_{it} + CD_i r_{it} + f(P,\zeta)) \\ &= \begin{cases} \min\limits_{\lambda\geq 0,\vartheta\in R} \lambda\cdot d_w + \dfrac{1}{N}\sum\limits_{k\in N}\vartheta_k \\ s.t f(\zeta) + \lambda(\zeta - \zeta_k) \leq \vartheta_k,\ \forall k \leq N \\ f(\bar{\zeta}) - \lambda(\bar{\zeta} - \zeta_k) \leq \vartheta_k,\ \forall k \leq N \\ f(\zeta_k) \leq \vartheta_k,\ \forall k \leq N \end{cases} \end{aligned} \tag{31}$$

Since the computational complexity of (31) is positively correlated with the size of samples, the following approximate equivalence is adopted to improve the computational efficiency. For more details of the model reformulation please see Appendix A, which provides a comprehensive derivation process.

$$\begin{cases} \inf\limits_{\lambda\geq 0,\vartheta\in R} \lambda\cdot d_w + \dfrac{1}{N}\sum\limits_{k\in N} f(\zeta_k) \\ s.t f'(\bar{\zeta}) \leq \lambda \\ -f'(\zeta) \leq \lambda \\ \lambda = \max(f'(\bar{\zeta}), -f'(\zeta)) \end{cases} \tag{32}$$

IV. COMPUTATIONAL RESULTS

A. PARAMETER SETTING

Numerical calculations were carried out on the Case-118 and Case-1888rte systems from MATPOWER 6.0, they represent the IEEE 118-bus transmission system and the French very high voltage transmission grid with 1888 buses, respectively [28]. Four wind farms rated at 400 MW are connected to buses 12, 49, 59 and 89 for Case-118. Six wind farms rated at 400 MW are connected to buses 355, 707, 921, 1628, 1651 and 1785 for Case-1888rte. The prediction errors, predicted and actual values of wind power are derived from TENNET [29], since the values are too large, they are reduced by 100 times for calculation. Load data and generator parameters can be obtained from MATPOWER 6.0 toolbox. Half, eighty and forty percent of the linear term coefficients of the quadratic cost curves are taken as the reserve, start-up and shut-down costs of thermal units, respectively. Fifty outliers are generated from a continuous uniform distribution over the interval [25], [29]. Since the data of wind farms are concentrated in [−39, 25], the values in [25] and [29] are suitable to be used as outliers. Given that the normal sample size is above 1920 and there are usually not too many outliers, we set the number of outliers to fifty. The proposed model was solved by the Cplex solver through GAMS.

TABLE 1. Technical details of three methods.

Methods	Distance Measure	Central Distribution	Support for Uncertainties
EDRO	Wasserstein	Empirical	Empirical
KDRO	Wasserstein	KDE-based	KDE-based
RDRO	Wasserstein	RKDE-based	RKDE-based

TABLE 2. Calculation time of RKDE processing time series.

Sample size	1920	3840	5760	7680	9600	11520
Calculation time (s)	5.06	10.24	64.18	51.56	90.69	289.57

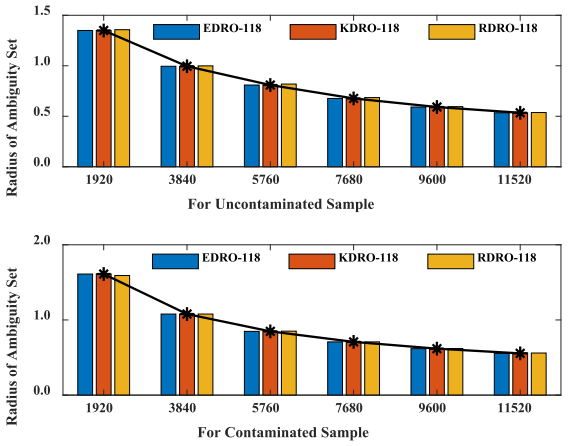

FIGURE 2. The radii of ambiguity set for both uncontaminated and contaminated samples in Case 118.

To verify the superiority of the proposed data-driven method in terms of statistical regularity and outlier suppression, we present tests on EDRO, KDRO and RDRO. The technical details of the above three methods and the calculation time of RKDE processing time series are shown in Table 1 and Table 2.

B. VERIFICATION OF STATISTICAL REGULARITY

As shown in Fig. 2 and Fig. 3, the radii of ambiguity set obtained by three methods tend to decrease with the expansion of sample size, which is true for both uncontaminated and contaminated samples.

This is because the empirical, KDE-based and RKDE-based distributions all tend to the real distribution with more available data at hand. So the radii gradually shrink.

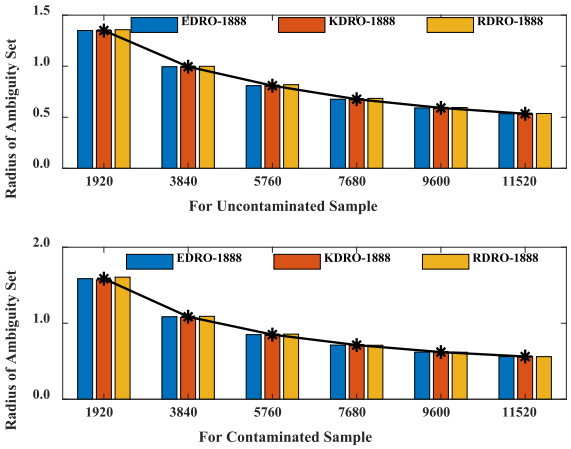

FIGURE 3. The radii of ambiguity set for both uncontaminated and contaminated samples in Case 1888rte.

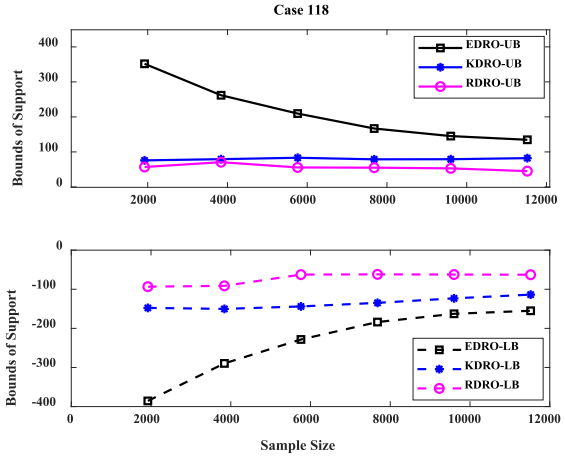

FIGURE 4. Upper bounds (UB) and lower bounds (LB) of the support space with increasing sample size for Case 118.

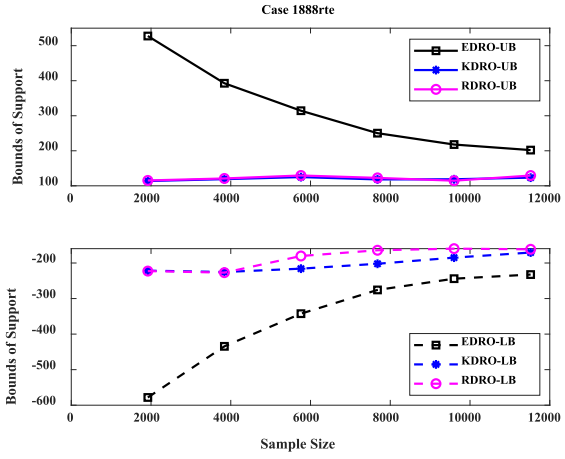

FIGURE 5. Upper bounds (UB) and lower bounds (LB) of the support space with increasing sample size for Case 1888rte.

Furthermore, due to the interference of abnormal data, the radii under contaminated samples will be larger, which will improve the conservatism of the model.

Fig. 4 and Fig. 5 depict the impacts of sample size on the bounds of the support space. Since the characteristics

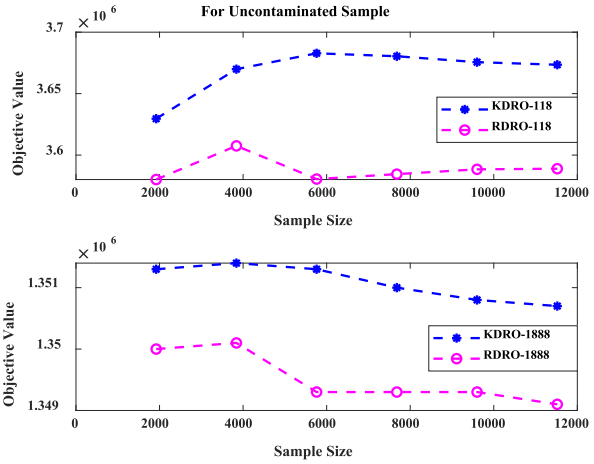

FIGURE 6. Operating costs of KDRO and RDRO with different sizes of the uncontaminated sample.

of the support space of contaminated samples are similar to those of the uncontaminated samples, only the results of the uncontaminated samples are analyzed here. As depicted in Figs. 4 and 5, the upper bounds (UB) of support by the three methods have a decreasing trend with increasing sample size. The lower bounds (LB) show an opposite trend. It indicates that the support spaces of the three methods shrink with the increasing sample size, which conforms to the statistical laws. As the distributions by KDE and RKDE are closer to the real distribution, their support space is smaller than that of EDRO. Since RKDE allocates different samples with diverse weights, it is closer to the real distribution, so its support space is smaller than that of KDE.

C. VERIFICATION OF OUTLIER SUPPRESSION ABILITY

From the results in Section IV-B, we know that the results of RDRO and KDRO are superior to EDRO. In order to further verify the outlier suppression ability of the proposed approach, the results of RDRO and KDRO are discussed below.

The objective values obtained by KDRO and RDRO are shown in Fig. 6 and Fig. 7. As the sample size increases, the distributions derived by KDE and RKDE are close to the real distribution and the conservatism of the model declines, leading to a reduction in operating costs of both KDRO and RDRO. In addition, for the contaminated samples, due to the interference of erroneous data, the cost curve rises. However, there is a significant increase in the objective value for KDRO, while it is not obvious in RDRO due to its immunity to outliers.

In addition, it should be noted that for both KDE and RKDE, their parameters fluctuate with increasing data size, resulting in some fluctuations in the objective value, which is determined by the data-driven feature of these methods. However, with increasing data size, the objective value stabilizes.

Fig. 8 and Fig. 9 describe the support size of uncertainties for both uncontaminated and contaminated samples. Evidently, in either case, the support space of RDRO is smaller

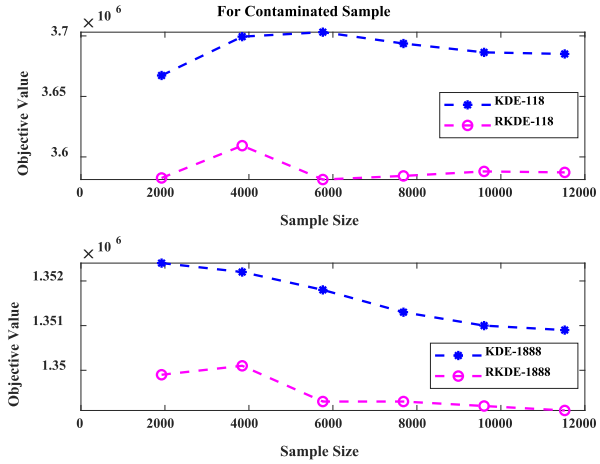

FIGURE 7. Operating costs of KDRO and RDRO with different sizes of the contaminated sample.

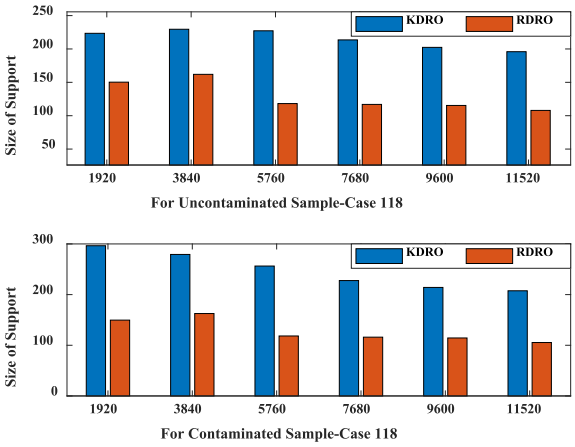

FIGURE 8. Support size of uncertainties for both uncontaminated and contaminated samples for Case 118.

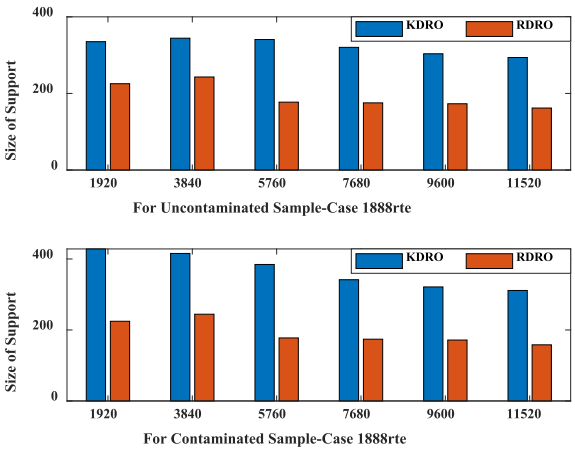

FIGURE 9. Support size of uncertainties for both uncontaminated and contaminated samples for Case 1888rte.

than that of KDRO, thus the conservatism is effectively reduced. More importantly, since it is affected by outliers, the support space by KDE will increase after the sample is contaminated, leading to excessive conservatism. But for RKDE, there is almost no change. The reason for this is

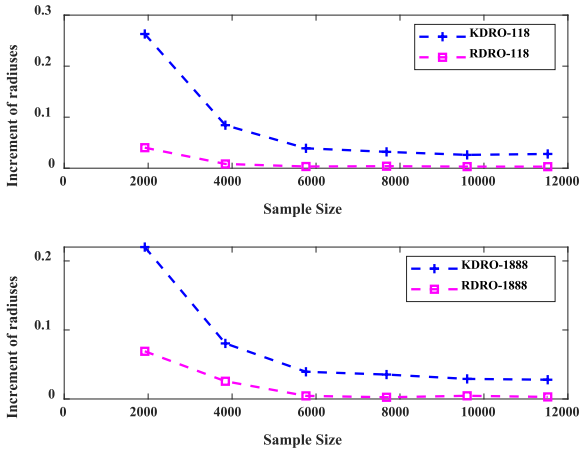

FIGURE 10. Radius increment of the contaminated sample relative to the uncontaminated sample by KDRO and RDRO.

TABLE 3. Calculation time of EDRO, KDRO and RDRO on case 118 for uncontaminated samples (US) and contaminated samples (CS).

Sample	EDRO-118		KDRO-118		RDRO-118	
size	US	CS	US	CS	US	CS
1920	8.2s	7.8s	7.5s	7.7s	7.4s	7.7s
3840	14.5s	14.5s	14.2s	14.5s	14.1s	14.3s
5760	21.1s	21.2s	21.0s	21.0s	20.8s	21.1s
7680	27.6s	27.7s	27.9s	27.8s	27.5s	27.8s
9600	34.4s	34.6s	34.3s	34.1s	34.3s	34.3s
11520	41.1s	41.2ss	41.3s	41.0s	40.9s	40.8s

TABLE 4. Calculation time of EDRO, KDRO and RDRO on case 1888rte for uncontaminated samples (US) and contaminated samples (CS).

Sample	EDRO-1888rte		KDRO-1888rte		RDRO-1888rte	
size	US	CS	US	CS	US	CS
1920	8.6s	8.4s	8.3s	8.2s	8.3s	8.3s
3840	12.4s	12.6s	12.3s	12.5s	12.3s	12.4s
5760	16.3s	16.6s	16.4s	16.5s	16.5s	16.5s
7680	20.3s	20.6s	20.4s	20.6s	20.6s	20.6s
9600	24.5s	24.7s	24.3s	24.6s	24.6s	26.1s
11520	28.5s	28.6s	28.5s	28.6s	28.4s	28.7s

that when deducing the probability distribution using RKDE, the outliers are assigned a very small weight, which has a minimal impact on the estimation of probability density.

The results shown in Fig. 10 are the radius increment of the ambiguity set of the contaminated sample relative to the uncontaminated sample by KDRO and RDRO. Due to the influence of outliers, the radii of the contaminated samples have an increment, but RDRO is less affected, so its increment is almost negligible. Although both KDE and RKDE methods converge to the real distribution with increasing data size, RKDE weakens the influence of outliers in the process of probability density estimation, so it has a stronger ability to suppress contaminated data.

Table 3 and Table 4 reveal the calculation time by EDRO, KDRO and RDRO on Case 118 and Case 1888rte. As the

sample size becomes larger, the amount of data processing increases, resulting in longer calculation times. It is worth noting that there is no significant difference in the calculation time among the three methods, but the numerical results of RDRO are relatively better, for both the uncontaminated and contaminated samples.

V. CONCLUSION

In the collection and transmission of renewable energy data, outliers are inevitable. Since empirical distributions are used in EDRO, the influence of outliers cannot be reduced. As for the KDE-based KDRO, since KDE gives the same weight to all data in the density estimation, it cannot weaken the influence of outliers either. The proposed data-driven robust approach integrates RKDE into distributionally robust optimization theory to estimate the central distribution. In this process, contaminated sample data are assigned very small weights, thus reducing their contribution to density estimation. Moreover, RKDE can converge to the true density quickly as more data are available, so that it can achieve robustness with low conservatism. By means of the calculations on case-118 and case-1888rte systems, the following conclusions can be drawn from the test results. Firstly, the proposed method is less conservative than EDRO and KDRO, so the radius of ambiguity set, support size and total scheduling costs are smaller. Secondly, the proposed method can effectively overcome the influence of outliers, make density estimation more accurate, and avoid the decision deviation caused by the contaminated sample.

APPENDIX
MODEL REFORMULATION PROCESS

The model reformation requires the following theorem:

$$\max_{F\in F_s} \mathrm{E}_F\{F(\zeta)\} = \min_{\lambda\geq 0}\left\{\lambda\cdot\varepsilon + \frac{1}{N}\sum_{k=1}^{N}\max_{\zeta\in\mathbb{S}} F(\zeta) - \lambda\left\|\zeta-\hat{\zeta}_k\right\|_1\right\} \tag{33}$$

The objective function $F(\bullet)$ can be transformed into a function with regard to wind power prediction errors.

$$\begin{aligned}
&\mathrm{E}_F(CU_i\bar{r}_{it} + CD_i r_{it} + f(P,\zeta))\\
&= \mathrm{E}_F(CU_i\bar{r}_{it} + CD_i r_{it} + a_i(P_{it}+\alpha_{it}\zeta_t)^2\\
&\quad + b_i(P_{it}+\alpha_{it}\zeta_t) + c_i))\\
&= \mathrm{E}_F(f(\zeta_t))\\
&= \begin{cases} f(\zeta_t) = a'\zeta_t^2 + b'\zeta_t + c' \\ a' = \sum_{t\in T}\sum_{i\in I} a_i\alpha_{it}^2 \\ b' = \sum_{t\in T}\sum_{i\in I} 2a_iP_{it}\alpha_{it} + b_i\alpha_{it} \\ c' = \sum_{t\in T}\sum_{i\in I}(a_iP_{it}^2 + b_iP_{it} + c_i) + CU_i\bar{r}_{it} + CD_ir_{it} \end{cases}
\end{aligned} \tag{34}$$

Combining the theorem in (33), (34) can be transformed into the following form:

$$\begin{aligned}
&\max_{F\in F_s} \mathrm{E}_F(f(\zeta_t))\\
&= \begin{cases} \min\limits_{\lambda\geq 0,\vartheta\in R} \lambda\cdot d_w + 1/N\sum\limits_{k\in N}\vartheta_k \\ s.t.\ \max\limits_{\lambda\geq 0,\vartheta\in R}(f(\zeta_t) - \lambda\cdot|\zeta-\zeta_k|) \leq \vartheta_k,\ \forall k\leq N \end{cases}\\
&= \begin{cases} \min\limits_{\lambda\geq 0,\vartheta\in R} \lambda\cdot d_w + 1/N\sum\limits_{k\in N}\vartheta_k \\ s.t. f(\zeta) + \lambda(\zeta-\zeta_k) \leq \vartheta_k,\ \forall k\leq N \\ f(\bar{\zeta}) - \lambda(\bar{\zeta}-\zeta_k) \leq \vartheta_k,\ \forall k\leq N \\ f(\zeta_k) \leq \vartheta_k,\ \forall k\leq N \end{cases}
\end{aligned} \tag{35}$$

In (35), the number of constraints is three times the number of samples, namely $3N$. In order to improve the calculation efficiency, we take $f(\zeta_k) = \vartheta_k$, then the following expression can be obtained:

$$\begin{cases} f(\zeta) + \lambda(\zeta-\zeta_k) \leq f(\zeta_k),\ \forall\zeta_k\in[\zeta,\bar{\zeta}] \\ f(\bar{\zeta}) - \lambda(\bar{\zeta}-\zeta_k) \leq f(\zeta_k),\ \forall\zeta_k\in[\zeta,\bar{\zeta}] \end{cases} \tag{36}$$

where

$$\begin{cases} \lambda \geq -\dfrac{f(\zeta_k)-f(\zeta)}{\zeta_k-\zeta} = -f'(\zeta) \\ \lambda \geq \dfrac{f(\bar{\zeta})-f(\zeta_k)}{\bar{\zeta}-\zeta_k} = f'(\bar{\zeta}) \end{cases} \tag{37}$$

Eventually, the approximate equivalent expression can be formed, as shown in (38), with a significantly reduced computational complexity.

$$\begin{cases} \inf\limits_{\lambda\geq 0,\vartheta\in R} \lambda\cdot d_w + \dfrac{1}{N}\sum\limits_{k\in N} f(\zeta_k) \\ s.t. f'(\bar{\zeta}) \leq \lambda \\ -f'(\zeta) \leq \lambda \\ \lambda = \max(f'(\bar{\zeta}), -f'(\zeta)) \end{cases} \tag{38}$$

REFERENCES

[1] A. Ma, J. Ji, and M. Khayatnezhad, "Risk-constrained non-probabilistic scheduling of coordinated power-to-gas conversion facility and natural gas storage in power and gas based energy systems," *Sustain. Energy, Grids Netw.*, vol. 26, Jun. 2021, Art. no. 100478.
[2] T. Ding, Z. Zeng, M. Qu, J. P. S. Catalão, and M. Shahidehpour, "Two-stage chance-constrained stochastic thermal unit commitment for optimal provision of virtual inertia in wind-storage systems," *IEEE Trans. Power Syst.*, vol. 36, no. 4, pp. 3520–3530, Jul. 2021.
[3] M. Malekpour, M. Zare, R. Azizipanah-Abarghooee, and V. Terzija, "Stochastic frequency constrained unit commitment incorporating virtual inertial response from variable speed wind turbines," *IET Gener., Transmiss. Distrib.*, vol. 14, no. 22, pp. 5193–5201, Nov. 2020.
[4] S. Naghdalian, T. Amraee, S. Kamali, and F. Capitanescu, "Stochastic network-constrained unit commitment to determine flexible ramp reserve for handling wind power and demand uncertainties," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4580–4591, Jul. 2020.
[5] Z. Chen, Z. Li, C. Guo, Y. Ding, and Y. He, "Two-stage chance-constrained unit commitment based on optimal wind power consumption point considering battery energy storage," *IET Gener., Transmiss. Distrib.*, vol. 14, no. 18, pp. 3738–3749, Sep. 2020.
[6] H. Qiu, W. Gu, W. Sheng, L. Wang, Q. Sun, and Z. Wu, "Resilience-oriented multistage scheduling for power grids considering nonanticipativity under tropical cyclones," *IEEE Trans. Power Syst.*, vol. 38, no. 4, pp. 3254–3267, Apr. 2023.

[7] L. Moretti, E. Martelli, and G. Manzolini, "An efficient robust optimization model for the unit commitment and dispatch of multi-energy systems and microgrids," *Appl. Energy*, vol. 261, Mar. 2020, Art. no. 113859.
[8] X. Zheng, H. Chen, Y. Xu, Z. Liang, and Y. Chen, "A hierarchical method for robust SCUC of multi-area power systems with novel uncertainty sets," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1364–1375, Mar. 2020.
[9] X. Yang, Z. Chen, X. Huang, R. Li, S. Xu, and C. Yang, "Robust capacity optimization methods for integrated energy systems considering demand response and thermal comfort," *Energy*, vol. 221, Apr. 2021, Art. no. 119727.
[10] Y. Zhang, X. Han, M. Yang, M. Wang, L. Zhang, P. Ye, and B. Xu, "Distributionally robust unit commitment based on imprecise Dirichlet model," *Proc. CSEE*, vol. 39, no. 17, pp. 5074–5084, 2019.
[11] B. Zhou, X. Ai, J. Fang, W. Yao, W. Zuo, Z. Chen, and J. Wen, "Data-adaptive robust unit commitment in the hybrid AC/DC power system," *Appl. Energy*, vol. 254, Nov. 2019, Art. no. 113784.
[12] K. Qu, X. Zheng, X. Li, C. Lv, and T. Yu, "Stochastic robust real-time power dispatch with wind uncertainty using difference-of-convexity optimization," *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4497–4511, Nov. 2022.
[13] C. Duan, L. Jiang, W. Fang, and J. Liu, "Data-driven affinely adjustable distributionally robust unit commitment," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1385–1398, Mar. 2018.
[14] M. B. Tookanlou, S. A. Pourmousavi, and M. Marzband, "Three-layer joint distributionally robust chance-constrained framework for optimal day-ahead scheduling of e-mobility ecosystem," 2021, *arXiv:2110.12123*.
[15] X. Fang, H. Cui, H. Yuan, J. Tan, and T. Jiang, "Distributionally-robust chance constrained and interval optimization for integrated electricity and natural gas systems optimal power flow with wind uncertainties," *Appl. Energy*, vol. 252, Oct. 2019, Art. no. 113420.
[16] C. Zhao and R. Jiang, "Distributionally robust contingency-constrained unit commitment," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 94–102, Jan. 2018.
[17] Y. Chen, Q. Guo, H. Sun, Z. Li, W. Wu, and Z. Li, "A distributionally robust optimization model for unit commitment based on Kullback–Leibler divergence," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5147–5160, May 2018.
[18] W. Hou, R. Zhu, H. Wei, and H. TranHoang, "Data-driven affinely adjustable distributionally robust framework for unit commitment based on Wasserstein metric," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 6, pp. 890–895, Mar. 2019.
[19] F. Sidun, C. Haozhong, and X. Guodong, "An extended quasi Monte Carlo probabilistic load flow method based on non-parametric kernel density estimation," *Autom. Electric Power Syst.*, vol. 39, no. 7, pp. 21–27, 2015.
[20] X. Xu, Z. Yan, M. Shahidehpour, Z. Li, M. Yan, and X. Kong, "Data-driven risk-averse two-stage optimal stochastic scheduling of energy and reserve with correlated wind power," *IEEE Trans. Sustain. Energy*, vol. 11, no. 1, pp. 436–447, Jan. 2020.
[21] Z. Wang, W. Wang, C. Liu, B. Wang, and S. Feng, "Short-term probabilistic forecasting for regional wind power using distance-weighted kernel density estimation," *IET Renew. Power Gener.*, vol. 12, no. 15, pp. 1725–1732, Nov. 2018.
[22] Y. Gao, Y. Sun, W. Yang, F. Xue, Y. Sun, H. Liang, and P. Li, "Study on load curve's classification based on nonparametric kernel density estimation and improved spectral multi-manifold clustering," *Power Syst. Technol.*, vol. 42, no. 5, pp. 1605–1612, 2018.
[23] L. Zeng, J. Xu, Y. Wang, Y. Liu, J. Tang, M. Wen, and Z. Chen, "Day-ahead interval scheduling strategy of power systems based on improved adaptive diffusion kernel density estimation," *Int. J. Electr. Power Energy Syst.*, vol. 147, May 2023, Art. no. 108850.
[24] C. Li, L. Zhang, Z. Ou, Q. Wang, D. Zhou, and J. Ma, "Robust model of electric vehicle charging station location considering renewable energy and storage equipment," *Energy*, vol. 238, Jan. 2022, Art. no. 121713.
[25] Q. Wang and L. Zhao, "Data-driven stochastic robust optimization of sustainable utility system," *Renew. Sustain. Energy Rev.*, vol. 188, Dec. 2023, Art. no. 113841.
[26] J. S. Kim and C. Scott, "Robust kernel density estimation," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2529–2565, 2012.
[27] C. Duan, W. Fang, L. Jiang, L. Yao, and J. Liu, "Distributionally robust chance-constrained approximate AC-OPF with Wasserstein metric," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4924–4936, Sep. 2018.
[28] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, Feb. 2011.
[29] *Tennet*. Accessed: Jul. 27, 2022. [Online]. Available: https://www.tennet.eu/?L=0#&panel1-1

WENTING HOU was born in Henan, China, in 1990. He received the B.S. degree in power system and automation from Hunan University, Changsha, China, in 2014, and the M.S. and Ph.D. degrees from Guangxi Key Laboratory of Power System Optimization and Energy Technology, Guangxi University, Guangxi, China, in 2016 and 2019, respectively.

He is currently with the School of Mechanical and Electrical Engineering, Zhoukou Normal University, Henan. He holds one book and two patents. His research interests include power system and integrated energy system optimization, distributionally robust optimization, data-driven approaches, and kernel density estimation.

LONGXIAN YI was born in Guangxi, China, in 1989. He received the dual B.S. degree in electrical engineering and automation and computer and application from Southwest University and Guangxi University, in 2018. He is currently pursuing the M.S. degree with Guangxi University.

Since 2020, he has been an Electrical Engineer with Nanning Branch of Guangxi Radio and Television Technology Center and a Senior Information System Project Manager. He is an Expert in the evaluation of power engineering, information technology services, and high-voltage distribution equipment for Guangxi government procurement projects. His research interests include power system computation, power supply and distribution and its stability, and the security and transmission of broadcasting and television signals.

HUAIBIN MIAO was born in Zhoukou, Henan, China, in 1988. He received the B.S. degree in agricultural mechanization and automation from Jiamusi University, Jiamusi, China, in 2013, and the M.S. and Ph.D. degrees in agricultural mechanization engineering from Jilin University, Changchun, China, in 2020.

Since 2020, he is a Lecturer of mechanical engineering with Zhoukou Normal University, Henan. His research interests include engineering mathematical model analysis, biomechanical testing and analysis, and the design of exoskeleton robots.

Dr. Miao is a member of the International Society of Bionic Engineering.

YINING MA was born in Zhoukou, Henan, China, in 2000. He is currently pursuing the B.S. degree with the School of Mechanical Engineering, University of Leeds, U.K.

His research interests include automatic engineering, automatic control, and engineering mathematical model analysis.