Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Retrieval-Augmented Non-Parametric Probabilistic Wind Power Forecasting under Regime-Dependent Uncertainty

Hongbin Sun[1]

dq_shb@ccit.edu.cn

Changchun Institute of Technology

Shijun Guo[1]

Changchun Institute of Technology

Xin Pan[2]

School of Computer Technology and Engineering

Zhanpeng Qu[1]

Changchun Institute of Technology

Jianchuan Ma[1]

Changchun Institute of Technology

Yaqi Xu[1]

Changchun Institute of Technology

Article

Keywords: wind power forecasting, probabilistic forecasting, non-parametric estimation, historical regime retrieval, multi-source weighting, interpretability

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11387551/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Retrieval-Augmented Non-Parametric Probabilistic Wind Power Forecasting under Regime-Dependent Uncertainty

Hongbin Sun[1]*, Shijun Guo[1], Xin Pan[2], Zhanpeng Qu[1], Jianchuan Ma[1] and Yaqi Xu[1]

[1] School of Electrical and Information Engineering, Changchun Institute of Technology, Changchun 130012, China
[2] School of Computer Technology and Engineering, Changchun Institute of Technology, Changchun 130012, China
* Correspondence: dq_shb@ccit.edu.cn

Reliable ultra-short-term probabilistic wind-power forecasting is important for renewable-energy integration, especially during ramps and extreme fluctuations. Existing deep probabilistic models typically encode historical operating regimes only implicitly, while conventional non-parametric estimators rely on low-dimensional similarity measures and coarse sample weighting. This paper proposes a retrieval-augmented non-parametric framework that reconstructs the conditional distribution of wind power from physically relevant historical analogs. A physically informed BiLSTM–Attention network extracts regime features from meteorological and historical power inputs; candidate historical samples are then retrieved and weighted by combining deep-feature similarity, wind-speed proximity, and temporal continuity, and the weighted neighbors are passed to a diffusion-smoothed kernel density estimator that produces both point forecasts and prediction intervals. Experiments on two heterogeneous wind farms in Zhejiang and Inner Mongolia show that, on Dataset A at the 60-min horizon, the method lowers MAE and RMSE by 21.2% and 22.7% relative to a Transformer baseline and reduces the 80% Winkler score by 24.4% relative to CKDE. A retrieved-neighbor analysis further makes each forecast traceable to physically meaningful historical samples.

Keywords: wind power forecasting; probabilistic forecasting; non-parametric estimation; historical regime retrieval; multi-source weighting; interpretability.

1. Introduction

Driven by the global energy transition and carbon-neutrality targets, the installed capacity of wind power has expanded rapidly. Wind power output is jointly modulated by wind speed, temperature, pressure, and other meteorological factors, and therefore exhibits pronounced randomness, volatility, and intermittency, posing acute challenges to power-system security and dispatch [1][2]. On the ultra-short-term horizon (15 min to 1 h), accurate wind power forecasting underpins real-time dispatch, automatic generation control (AGC), energy-storage scheduling, and electricity-market clearing. Compared with deterministic point forecasting [3], probabilistic forecasting reports the predictive distribution or its intervals and thus quantifies risk for downstream decisions, and has consequently attracted growing research attention [4].

Wind power forecasting methods have evolved from classical statistical models to machine-learning and deep-learning approaches. Persistence models, ARIMA, SVR, random forests, LSTM, GRU, CNN, and Transformer-type architectures have improved deterministic forecasting accuracy [5]–[11]. However, most of these methods are designed as point predictors and cannot directly characterize predictive uncertainty, which motivates the development of probabilistic forecasting methods. Existing probabilistic approaches can be broadly grouped into parametric, end-to-end deep probabilistic, and non-parametric methods.

The first category is parametric: a distribution family is postulated and its parameters are learned. Deng et al. [12] introduced prior distributions over LSTM weights to construct a Bayesian LSTM, and Chen et al. [13] proposed a quantile-regression LSTM. Such methods are easy to implement and computationally

efficient, but their expressiveness is bounded by the assumed parametric form [14]. Wind power exhibits regime-dependent distributional shapes [15]: a single parametric distribution can rarely capture, simultaneously, the truncation in the low-speed regime, the skewness or multi-modality in the mid-speed regime, and the concentration in the high-speed regime [16].

The second category comprises end-to-end deep probabilistic models, which directly output distributions, quantiles, or scenarios. Liu et al. [17]developed a Bayesian BiGRU framework for ultra-short-term forecasting; Wen et al. [18] used conditional normalizing flows for distribution-free probabilistic forecasting; Zheng et al. [19] modeled regime-dependent stochasticity with conditional variational auto-encoders; and Salinas et al. [20] introduced DeepAR for autoregressive probabilistic time-series forecasting. Wang et al. [21]proposed a deep asymmetric Laplace network for joint deterministic and probabilistic forecasting, while Zhang et al. [22] designed a multi-source temporal attention network. However, these methods encode historical regimes mainly through network parameters, without explicitly identifying historically analogous operating states. Consequently, under abrupt regime shifts, they may fail to surface relevant historical evidence in time, limiting generalization [23] [24].

The third category comprises non-parametric methods based on kernel density estimation (KDE), which smooth historical samples through a kernel without postulating a parametric form and offer transparent physical meaning [25]. Jeon and Taylor [26] applied conditional KDE to wind-power density forecasting, and Wang et al.[27] developed a distance-weighted KDE that emphasizes locally similar samples. Beyond these, Gu et al. [28] combined LSTM with non-parametric KDE for short-term forecasting and uncertainty analysis; Jeon et al. [29] discussed forecast reconciliation in probabilistic forecasting of wind power and electric load; and Wang et al. [30] addressed wind-power-curve modeling under inconsistent data. Despite these advances, existing weighting schemes remain limited: vanilla KDE treats all historical samples equally and cannot differentiate sample relevance across regimes, while previous weighting refinements typically rely on a single conditioning variable or a simplistic distance metric and fail to capture the multi-dimensional nature of regime similarity. Fusing multi-source information within the KDE framework to enable refined retrieval and dynamic weighting of historical samples therefore remains an open problem for probabilistic wind-power forecasting.

Closely related to KDE-based methods is the analog ensemble (AnEn) paradigm, which also constructs probabilistic forecasts from historically similar samples [31]. Alessandrini et al. [32] applied AnEn to short-term probabilistic wind-power forecasting, and Cervone et al. [33] further coupled AnEn with neural networks. However, classical AnEn usually relies on hand-crafted distances over low-dimensional meteorological variables and often treats analogs with limited or equal weighting [34]. This makes it difficult to capture high-dimensional nonlinear regime similarity and motivates a more refined retrieval and weighting mechanism.

In summary, existing probabilistic wind-power forecasting methods remain limited in their explicit use of historically analogous regimes. Parametric and end-to-end deep models encode historical information mainly through model parameters, whereas KDE- and AnEn-type methods use historical samples more directly but often rely on low-dimensional similarity metrics and coarse weighting schemes.

To address this gap, this paper formulates ultra-short-term probabilistic wind-power forecasting as an explicit historical-regime retrieval problem. A physically informed BiLSTM–Attention model learns regime representations, while deep-feature similarity, wind-speed proximity, and temporal continuity are

jointly used to retrieve and weight historical analogs. The weighted samples are then used in a diffusion-smoothed KDE to reconstruct the conditional wind-power distribution and generate prediction intervals. This design preserves the flexibility of non-parametric density estimation while making each forecast traceable to specific historical samples. The main contributions are threefold:

1. An explicit regime-retrieval framework is proposed for ultra-short-term probabilistic wind-power forecasting, enabling the current operating state to be linked with historically analogous regimes rather than being modeled only implicitly.

2. A multi-source dynamic weighting strategy is developed to combine deep-feature similarity, physical wind-speed proximity, and temporal continuity, allowing historical samples to be selected and weighted more adaptively for conditional density estimation.

3. The framework is evaluated under normal, ramp, and extreme-fluctuation regimes, and retrieved-neighbor analysis links individual forecasts to specific historical samples.

2. Methods

2.1 Overall framework

The proposed retrieval-augmented non-parametric probabilistic forecasting framework is illustrated in Fig. 1. It consists of three main modules. First, a physically informed BiLSTM–Attention network extracts the current regime representation and provides an initial point forecast. Second, historical samples are retrieved and weighted using deep-feature similarity, physical wind-speed distance, and temporal continuity. Third, a diffusion-smoothed weighted KDE reconstructs the conditional power distribution and generates quantiles and prediction intervals. The final point forecast is obtained by combining the deep-network output with the KDE mean, with key hyper-parameters tuned by PSO.

The input comprises 24 consecutive time steps at a 15-min resolution and 35 engineered features, as summarized in Fig. 1.

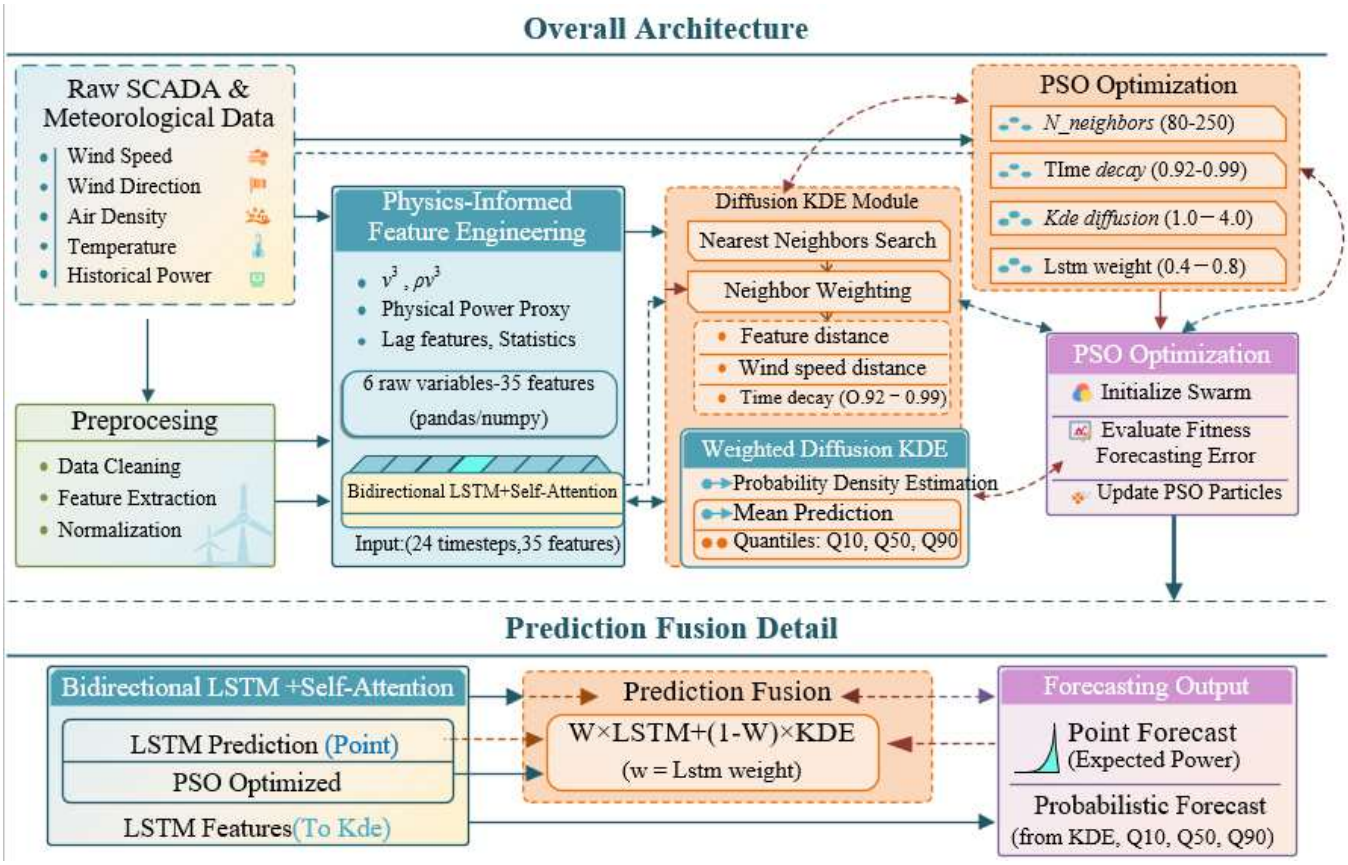


Fig. 1 Overall framework of the proposed multi-source weighted probabilistic forecasting method.

2.2 Data and preprocessing

To systematically evaluate the proposed method under contrasting operating environments, two real wind-farm datasets with substantial regional differences are used. Dataset A originates from a wind farm in Zhejiang Province, China, and Dataset B from a wind farm in Inner Mongolia. The two regions differ markedly in climatic background, wind resource, and weather evolution. Zhejiang represents a humid coastal climate strongly influenced by sea–land interaction, monsoon activity, and local weather processes; meteorological conditions vary frequently and the wind power output exhibits strong short-term variability and non-stationarity. Inner Mongolia is dominated by a continental climate with stronger and steadier wind resources, drier air, and larger diurnal temperature ranges, with weather systems and wind-speed distributions that differ substantially from those over the coast.

Table 1 Characteristics of the two wind-farm datasets.

Statistic	Dataset A	Dataset B
Region	Zhejiang	Inner Mongolia
Climate	Humid coastal	Continental
Sample size	61 190	70 176
Installed capacity (MW)	245	201.5
Mean power (MW)	72.76	72.68
Std. of power (MW)	80.39	55.72
Mean wind speed (m/s)	4.76	6.30
Wind-speed range (m/s)	0.00 – 25.44	0.00 – 23.30
Time span	2021-01-01 – 2022-09-30	2019-01-01 – 2020-12-31

Dataset A is split chronologically into training (70%), validation (15%) and test (15%) subsets for network training, parameter selection and in-domain evaluation, respectively. Standardization parameters are estimated on the training set and applied unchanged to validation and test data. Parameters are selected on the validation set, and no choice depends on the test set. At test time, retrieval uses the training archive exclusively; the archive precedes the held-out samples in Dataset A. Each forecasting horizon is modeled directly and independently. Dataset B is excluded from training and parameter selection and is used as a fixed external dataset to evaluate transfer between wind farms. This cross-regional test evaluates transfer across the two historical records, rather than a chronological deployment from Dataset A to Dataset B.

The raw data contain sensor faults, communication outages, and recording errors and therefore require cleaning. Because Dataset A is used for both training and the main comparative experiments, its preprocessing is described in detail; the same principles are applied to Dataset B in the generalization experiments to ensure consistency and comparability.

Cleaning rules tailored to wind-farm data are adopted: samples missing either power or wind speed are removed; wind directions outside [0°, 360°) are corrected by modular arithmetic; and physically implausible pressure or air-density records, identified according to the data unit and valid sensor range, are replaced with the column median rather than the mean to reduce sensitivity to extreme sensor errors. Because the present study targets the aggregate output of a wind farm rather than that of a single turbine, gradual decreases in farm-level output during cut-out under high winds (with turbines progressively triggering protection) reflect normal operation and must be preserved — a single-turbine cut-in/cut-out filter is therefore inappropriate. Two additional categories of abnormal samples are then removed: (i) low-wind-speed but high-power outliers, where wind speed is below the cut-in (3 m/s) but power exceeds 50 MW, indicating insufficient meteorological-mast representativeness or temporal misalignment; and (ii) farm-wide shutdowns, where wind speed exceeds 8 m/s but the farm output is below 5 MW, reflecting curtailment, protection actions, or scheduled maintenance. Supplementary Fig. S1 compares the power–wind-speed scatter before and after cleaning of Dataset A.

2.3 Construction of physical features

From first principles of wind-energy conversion, the mechanical power of a wind turbine is given by

$$P = \frac{1}{2} C_p \rho A v^3 \leftarrow \tag{1}$$

where ρ is air density, A the swept rotor area, v the wind speed, and C_p the power coefficient. Guided by this relation, the following physical features are constructed and used as auxiliary network inputs.

The cubic wind-speed feature, which directly reflects the theoretically available wind energy:

$$f_1 = v^3 \leftarrow \tag{2}$$

An air-density–wind-speed coupling feature:

$$f_2 = \rho v^3 \tag{3}$$

with air density estimated by the ideal-gas equation of state:

$$\rho = p/(R_d T)^{\leftarrow} \tag{4}$$

Turbulence intensity, which characterizes the variability of wind speed:

$$f_3 = \sigma_v / \bar{v} \tag{5}$$

In addition, sine–cosine encoded periodic time features, lagged features of historical power and wind speed, and rolling-window statistics are introduced. All features are standardized before being fed into the network.

2.4 Deep temporal feature extractor

A bidirectional LSTM (BiLSTM) is adopted as the temporal backbone. BiLSTM processes the input sequence both forward and backward, capturing bidirectional temporal dependencies. The network is configured as a three-layer stack with a hidden size of 256 per direction and a dropout of 0.3 to mitigate overfitting. At each time step t, the BiLSTM produces the concatenation of the forward and backward hidden states:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \in R^{2d_h} \tag{6}$$

where d_h denotes the unidirectional hidden dimension.

To dynamically focus on the time steps most informative for forecasting, additive attention is applied over the BiLSTM outputs:

$$e_t = \mathrm{v}^T \tanh(\mathrm{W}_a h_t + \mathrm{b}_a) \tag{7}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)} \tag{8}$$

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{9}$$

where W_a, v and b_a are learnable parameters, α_t is the attention weight at time step t, and c is the context vector.

To preserve information from the raw input and alleviate information loss in deeper layers, a residual connection is introduced: the original input at the last time step x_T is linearly projected and concatenated with the attention context vector,

$$f = [c; \mathrm{W}_p x_T + \mathrm{b}_p] \tag{10}$$

with W_p the projection matrix. The concatenated vector is fed through a fully-connected layer to produce the point forecast y, and f simultaneously serves as the deep regime representation used downstream for neighbor retrieval.

Because forecasting errors during high-power periods exert a disproportionate impact on power-system operation, a power-related weighted MSE loss is designed to moderately up-weight high-power samples:

$$\mathrm{L} = \frac{1}{N}\sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2 \tag{11}$$

$$w_i = 1 + 0.3 \cdot \frac{y_i}{P_{max}} \tag{12}$$

where P_{max} is the maximum power in the training set. The design constrains the per-sample weight to [1.0, 1.3], up-weighting high-power samples without inducing the training instability that excessive weighting would cause.

2.5 Multi-source dynamic weighting strategy

Deep-feature similarity describes the learned operating regime, wind-speed proximity provides a physical anchor, and temporal proximity accounts for continuity between operating states. These three sources jointly determine the retrieved historical neighbors and their weights.

(1) Deep-feature similarity: Using the deep representation f extracted by the BiLSTM–Attention network of Section 2.4, the Euclidean distance between the current instant t and a historical instant t_i in the deep-feature space is

$$d_{feat}(t, t_i) = \| f_t - f_{t_i} \|_2 \tag{13}$$

(2) Physical wind-speed distance. Given the direct physical link between wind speed and power, the absolute distance in wind-speed space is defined as

$$d_{wind}(t, t_i) = | v_t - v_{t_i} | \tag{14}$$

(3) Temporal decay factor. Atmospheric states are temporally continuous, so historical samples closer in time to the current instant are typically of higher reference value. An exponentially decaying temporal weight is adopted:

$$w_{time}(t, t_i) = \lambda^{| t - t_i | / \tau} \tag{15}$$

where $\lambda \in (0, 1)$ is the decay coefficient and $\tau = 168$ is the decay period, corresponding to 42 h at the 15-min sampling resolution.

Because the three quantities differ in scale, the deep-feature distance and the wind-speed distance are first min–max normalized to [0, 1]:

$$\tilde{d}_{feat} = \frac{d_{feat} - d_{feat}^{min}}{d_{feat}^{max} - d_{feat}^{min}} \tag{16}$$

$$\tilde{d}_{wind} = \frac{d_{wind} - d_{wind}^{min}}{d_{wind}^{max} - d_{wind}^{min}} \tag{17}$$

The normalized deep-feature and wind-speed distances are combined with equal weight (α = 0.5), selected on the validation set. They describe similarity in the learned regime representation and physical proximity along the power curve, respectively:

$$d_{combined} = 0.5 \cdot \tilde{d}_{feat} + 0.5 \cdot \tilde{d}_{wind} \tag{18}$$

Dividing the combined distance by the temporal decay weight yields the composite score s: samples with larger temporal weights (i.e., closer in time) obtain smaller scores and are therefore more likely to be selected as neighbors:

$$s(t, t_i) = \frac{d_{combined}(t, t_i)}{w_{time}(t, t_i) + \grave{o}} \tag{19}$$

with $\varepsilon = 10^{-10}$ a numerical-stability term.

All historical samples are sorted by the composite score s in ascending order, and the K samples with the smallest scores are selected as the neighbor set. The selected neighbors are then weighted by an exponential transformation of their scores:

$$w_i = \exp(-2.\mathrm{s}_i) \tag{20}$$

This exponential mapping amplifies the contribution of neighbors with small composite scores (i.e., highly similar regimes), thereby strongly discriminating among regime similarities.

2.6 Diffusion-smoothed weighted KDE

Given the k weighted neighbors $\{(p_i, w_i)\}$, where p_i denotes the actual power of the $i-th$ neighbor, a weighted KDE is constructed to obtain the conditional density:

$$\hat{f}(x) = \frac{1}{\sum_{i=1}^{K} w_i} \sum_{i=1}^{K} w_i \cdot \frac{1}{h} \phi\left(\frac{x - p_i}{h}\right) \tag{21}$$

where $\phi(\cdot)$ is the standard Gaussian kernel and h is the bandwidth, initialized via Silverman's rule of thumb,

$$h = 1.06 \cdot \sigma_p \cdot K^{-1/5} \tag{22}$$

with σ_p the weighted standard deviation of the neighbors' power values.

Conventional KDE may exhibit local artifacts, especially when the sample size is limited or the weights are unevenly distributed. A diffusion equation is therefore applied to smooth the density estimate. Treating f(x) as the initial condition, smoothing is realized by solving

$$\frac{\partial u(x,t)}{\partial t} = D \frac{\partial^2 u(x,t)}{\partial x^2} \tag{23}$$

where $u(x,0)$ = f(x) is the initial density, D is the diffusion coefficient, and t is the diffusion time. A Crank–Nicolson implicit scheme is used, ensuring numerical stability while producing the smoothed density $u(x,t_{diff})$. The diffusion time t_{diff} is treated as a hyper-parameter optimized by PSO.

From the smoothed conditional density, we extract: (i) the KDE mean prediction, defined as the expected value of the smoothed density; (ii) the quantiles $q_{0.1}$ $q_{0.5}$, and $q_{0.9}$ obtained from the inverse CDF; and (iii) the 80% and 90% prediction intervals $[q_{0.1}, q_{0.9}]$ and $[q_{0.05}, q_{0.95}]$, respectively.

2.7 Hybrid prediction and parameter selection

The final point forecast is a weighted combination of the BiLSTM-based point prediction and the KDE mean:

$$\hat{y}_{final} = \omega \cdot \hat{y}_{LSTM} + (1-\omega) \cdot \hat{y}_{KDE} \tag{24}$$

where $\omega \in [0.4, 0.8]$ is the fusion weight, treated as a hyper-parameter optimized by PSO. This hybrid strategy combines the non-linear fitting capacity of the deep network with the locally adaptive estimation that KDE provides on analogous regimes.

2.8 Experimental settings and evaluation

The proposed method is compared with persistence, SVR[35], LSTM[7], GRU[8], Transformer[36], Informer[37], QR-LSTM[38], DeepAR[20], and CKDE[27]. Among these, Informer is a representative Transformer variant tailored for long-sequence forecasting; QR-LSTM and DeepAR represent deep probabilistic temporal models; CKDE represents the same family of non-parametric probabilistic methods. The point forecast of the proposed method is the hybrid forecast defined in Section 2.7, obtained by combining the BiLSTM-based point prediction with the KDE mean. To ensure a fair comparison, all deep-learning baselines share the identical preprocessing, the chronological 70%/15%/15% split, and the early-stopping protocol used by the proposed method.

For CKDE the conditioning variables are extended to (wind speed, lagged power, hour) in line with the standard configuration of modern non-parametric baselines, and its conditional bandwidth and number of neighbors are tuned on the validation set. We acknowledge that the proposed method retrieves analogs in

a deep-feature space whose dimension is far higher than the three conditioning variables of CKDE; this difference in conditioning capacity is itself part of the methodological contribution. To isolate the effect of the multi-source weighting from that of the deep representation, Section 3.6 presents a controlled comparison against a Deep-Feature AnEn (DF-AnEn) variant that shares our deep features but uses an equally-weighted ensemble, thereby attributing the gains to deep representation and to multi-source weighting in turn. Each deep baseline takes the raw meteorological and power features as input, without using the physical features constructed in Section 2.3, so as to fairly compare the "generic deep-learning framework" with the proposed "physical features + multi-source weighting" framework. Forecasting horizons are 15, 30, 45, and 60 minutes.

A combined evaluation system covering point and probabilistic forecasting is adopted.

Point-forecast metrics include mean absolute error (MAE), root-mean-square error (RMSE), and the coefficient of determination (R^2).

Probabilistic-forecast metrics include the prediction-interval coverage probability (PICP), the prediction-interval normalized average width (PINAW) at the 80% and 90% nominal levels, and the Winkler score.

The definitions of MAE, RMSE, R^2, PICP, PINAW and the Winkler score are given in the Supplementary Methods.

Code availability

The code used in this study is available from the corresponding author upon reasonable request.

3. Results

3.1 Point-forecasting performance

Table 2 reports MAE, RMSE and R^2 at the 15- and 60-min horizons on Dataset A; the complete four-horizon results are provided in Supplementary Table S7. Smaller MAE/RMSE and larger R^2 indicate better point-forecast performance.

At the 1-step horizon, it obtains the lowest MAE and RMSE among all methods. At the 4-step horizon, it reduces MAE and RMSE by 21.2% and 22.7%, respectively, relative to the Transformer baseline, while maintaining the highest R^2. Compared with Informer, the proposed method shows more stable RMSE and R^2 performance as the forecasting horizon increases, indicating better longer-horizon stability.

Under matched preprocessing and validation settings, the proposed method reduces MAE relative to CKDE by 47.0%, 27.5%, 30.1%, and 24.3% across the four horizons. This suggests that deep regime representation and multi-source neighbor weighting enable a more effective use of historical samples than low-dimensional conditional KDE.

Supplementary Fig. S2 displays the point-forecast metrics across the evaluated methods.

Table 2 Point forecasting performance at the 15- and 60-min horizons on Dataset A

Method	Horizon (min)	MAE (MW)	RMSE (MW)	R^2
Persistence	15	5.94	10.65	0.982
LSTM	15	8.89	13.89	0.977
Transformer	15	8.98	14.27	0.977
SVR	15	7.64	11.34	0.985
GRU	15	9.01	14.09	0.977
QR-LSTM	15	9.61	15.00	0.974
DeepAR	15	8.67	14.36	0.976
CKDE	15	9.80	15.44	0.973
Informer	15	5.97	9.29	0.990
Ours	15	5.19	8.54	0.992
Persistence	60	14.74	25.44	0.899
LSTM	60	19.47	29.46	0.901
Transformer	60	17.28	27.78	0.911
SVR	60	17.12	25.34	0.927
GRU	60	18.91	28.15	0.909
QR-LSTM	60	20.78	33.51	0.871
DeepAR	60	17.97	28.07	0.910
CKDE	60	17.99	26.82	0.918
Informer	60	13.57	22.51	0.942
Ours	60	13.61	21.46	0.947

3.2 Probabilistic-forecasting performance

The proposed method is compared with QR-LSTM, DeepAR, and CKDE — the three baselines of Section 3.1 that produce probabilistic outputs. Table 3 reports PICP, PINAW, and Winkler scores at the 80% and 90% nominal levels for all horizons. Supplementary Fig. S3 summarizes the 80% PICP, PINAW, and Winkler trends across forecasting horizons.

Table 3 shows that the proposed method achieves the best overall reliability–sharpness trade-off across the two confidence levels. At the 80% confidence level, $PICP_{80}$ ranges from 77.55% to 83.45%, remaining close to the nominal level across most horizons, while $PINAW_{80}$ and $Winkler_{80}$ are consistently lower than those of the baselines. At the 90% confidence level, $PICP_{90}$ ranges from 87.40% to 90.70%. Although the coverage is slightly below the nominal level at the 4-step horizon, the proposed method still produces much narrower intervals and lower $Winkler_{90}$ scores than the competing methods, indicating a more favorable balance between reliability and sharpness.

Table 3 Probabilistic forecasting performance at the 80% and 90% nominal levels on Dataset A

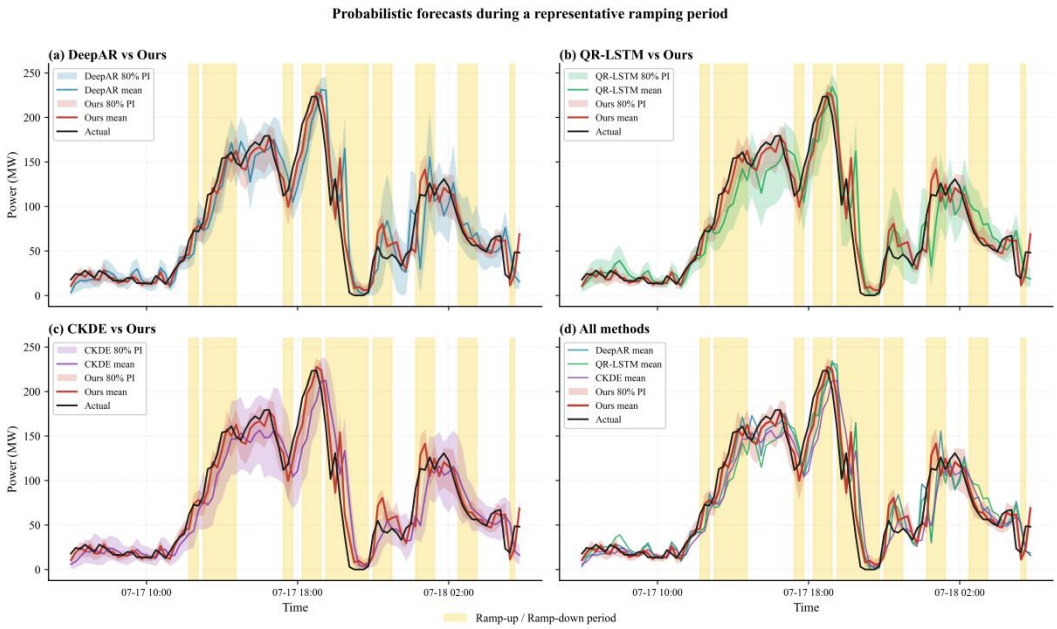
Method	Horizon (min)	$PICP_{80}$ (%)	$PINAW_{80}$	$Winkler_{80}$ (MW)	$PICP_{90}$ (%)	$PINAW_{90}$	$Winkler_{90}$ (MW)
QR-LSTM	15	81.25	0.131	46.20	91.80	0.186	55.02
DeepAR	15	83.85	0.110	43.01	89.35	0.139	55.80
CKDE	15	73.40	0.141	45.27	80.70	0.187	56.30
Ours	15	83.45	0.070	25.36	90.70	0.095	34.68
QR-LSTM	30	78.55	0.155	64.14	81.20	0.176	93.86
DeepAR	30	82.50	0.147	56.15	88.55	0.184	70.63
CKDE	30	71.75	0.176	57.88	80.60	0.237	71.98
Ours	30	82.05	0.110	40.79	88.05	0.134	53.99
QR-LSTM	45	69.00	0.165	88.70	76.45	0.201	130.93
DeepAR	45	79.05	0.167	69.79	87.00	0.210	89.29
CKDE	45	70.40	0.203	68.55	80.60	0.274	84.77
Ours	45	81.20	0.125	48.23	90.15	0.168	62.76
QR-LSTM	60	68.70	0.168	110.34	77.35	0.239	151.52
DeepAR	60	67.95	0.162	86.87	79.25	0.204	115.57
CKDE	60	68.40	0.228	77.89	80.20	0.310	97.24
Ours	60	77.55	0.137	58.87	87.40	0.180	74.06

3.3 Performance under non-stationary scenarios

Non-stationarity in wind power output manifests primarily as steep ramps and high-frequency fluctuations. Given the 245 MW installed capacity of Dataset A, the test set is partitioned as follows: (i) ramp events, with a power change $|\Delta P_{1h}|$ over the past hour exceeding $15\% \times C = 36.75$ MW, separated into ramp-up and ramp-down by direction; (ii) extreme-fluctuation events, with a one-hour standard deviation $\sigma_{1h} \geq 8\% \times C = 19.60$ MW; and (iii) normal regimes (the remaining samples). On each subset, the point accuracy (MAE) and interval quality ($PICP_{80}$, $PINAW_{80}$, Winkler) of the proposed method (Ours) and of three probabilistic baselines (DeepAR, QR-LSTM, CKDE) are computed. Results are reported in Supplementary Table S1 and Supplementary Fig. S4.

Under non-stationary regimes, the baselines show clear error amplification and coverage deterioration. Their MAE rises to 23.94–34.18 MW under ramp-up and 30.72–33.82 MW under extreme fluctuations, whereas the proposed method remains at 11.92–13.37 MW. Their $PICP_{80}$ also drops sharply, reaching only 49.5%–55.1% under ramp-up and 36.8%–57.9% under extreme fluctuations. In contrast, the proposed method achieves higher coverage than the baselines while producing narrower intervals, with the advantage being most evident under ramp-down and extreme-fluctuation regimes.

A representative ramp-intensive period is visualized in Fig. 2, covering the period from 09:00 on 17 July 2022 to 03:00 on 18 July 2022.



330 Fig. 2 80% prediction intervals during a ramp-intensive period (17 Jul 2022 09:00 – 18 Jul 2022 03:00)

In Fig. 2, the baseline mean forecasts lag the peaks and troughs during the ramp-intensive period, with intervals widening to compensate, whereas the proposed method follows rapid changes more closely and maintains sharper intervals. Together with the scenario-level results in Supplementary Table S1, this visualization indicates that the retrieval-based density reconstruction provides more reliable interval
335 estimates under non-stationary wind-power regimes.

3.4 Ablation study

To evaluate the contribution of the core components, four variants are compared: a Baseline consisting of BiLSTM–Attention and equally weighted diffusion-smoothed KDE; a +Weighting variant that adds the proposed multi-source dynamic weighting strategy; a +Physical variant that adds the constructed physical
340 features; and Ours, which combines both components. All variants follow the same data split, preprocessing, and hyper-parameter tuning protocol. Since the proposed method is mainly designed to improve probabilistic stability under more challenging forecasting conditions, the ablation study reports the 45-min and 60-min horizons, where error accumulation and uncertainty calibration are more pronounced. The results are shown in Supplementary Table S2 and Table 4.

345 The ablation results show that the two components do not provide independent additive gains. Adding either component alone slightly degrades point accuracy—at the 4-step horizon, MAE rises from 14.58 (Baseline) to 16.11 with weighting and to 15.66 with physical features—yet combining them yields the lowest MAE (13.61) and RMSE (21.46) across all variants. The benefit of multi-source weighting therefore depends on the regime representation used for neighbor retrieval.

Table 4 Ablation study at the 60-min horizon on Dataset A

Variant	MAE	RMSE	R^2	$PICP_{80}$	$PINAW_{80}$	$Winkler_{80}$
Baseline	14.58	22.52	0.941	87.05	0.187	65.23
+Weighting	16.11	26.12	0.922	72.50	0.125	76.71
+Physical	15.66	23.76	0.935	84.50	0.187	67.09
Ours	13.61	21.46	0.947	77.55	0.137	58.87

3.5 Retrieved-neighbor analysis

Four representative instants are selected from the test set: a ramp-up peak (18:00 on 17 July 2022), a ramp-down trough (22:00 on 17 July 2022), an extreme-fluctuation moment (21:00 on 7 July 2022), and a normal-regime control (12:00 on 13 July 2022). For each, the top-10 retrieved historical neighbors, their distance components, and the resulting weighted predictive density are recorded. Results are shown in Fig. 3

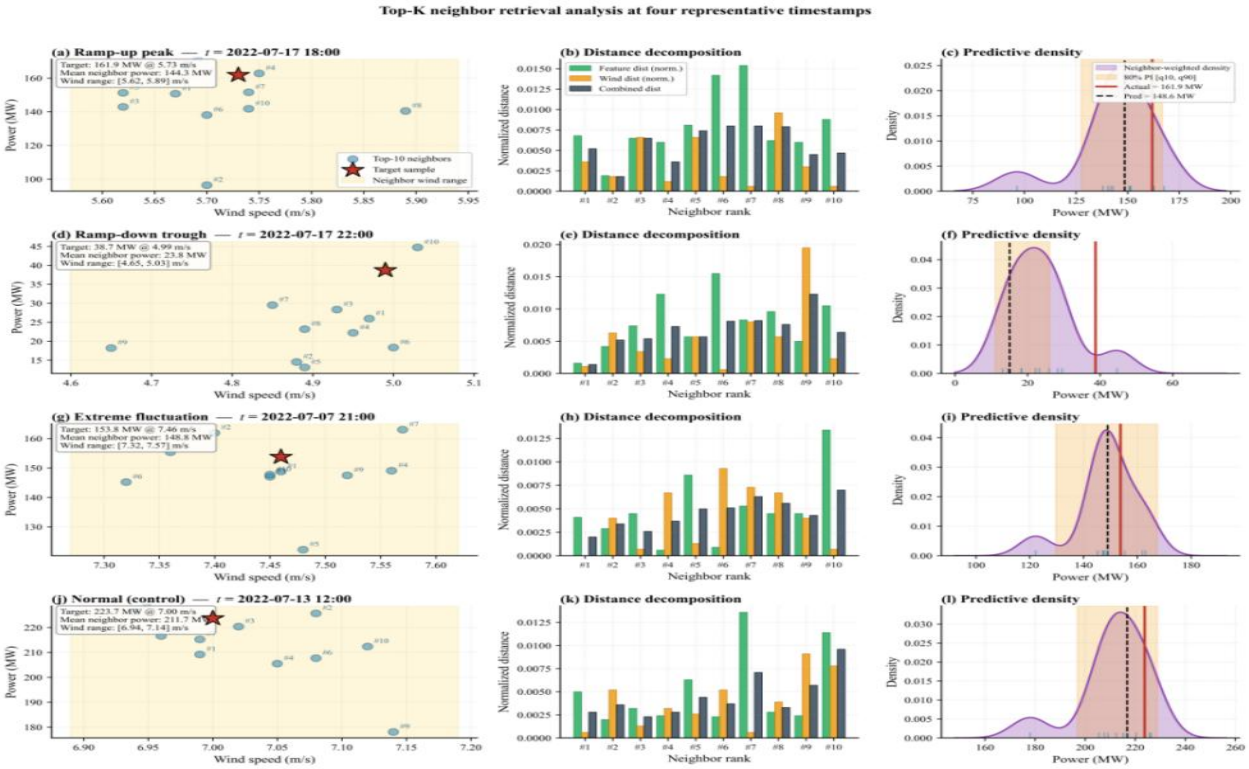


Fig. 3. Top-10 historical neighbors at four representative instants (scatter, distance decomposition, weighted predictive density).

Column 1 of Fig. 3 shows that the retrieved neighbors remain close to the target wind speed in all four representative regimes, with maximum deviations no larger than 0.16 m/s. This indicates that the physical wind-speed constraint keeps the selected samples within a plausible region of the power curve. Column 2 further shows that deep-feature distance and wind-speed distance contribute differently to the final ranking, confirming their complementary roles: wind-speed distance anchors retrieval to the local power-curve relationship, while deep-feature distance refines the ranking by capturing broader regime

characteristics. The retrieved neighbors are therefore selected by combined regime similarity rather than by a simple sliding-window rule.

A representative difficult case is shown in Fig. 3(d–f), where the true power is 38.71 MW, the point forecast is 15.05 MW, and the 80% prediction interval is [11.10, 26.07] MW. The retrieved neighbors mostly fall within a lower-power range, suggesting that comparable high-output samples are scarce in this local regime. Although the interval fails to cover this realization, the neighbor distribution provides a traceable explanation of the error and helps identify when the current condition departs from available historical analogs.

3.6 Comparison with the AnEn paradigm

The analog ensemble (AnEn) is a representative analog-based non-parametric forecasting paradigm and is closely related to the proposed retrieval framework. To clarify the difference, two comparators are designed. Classical AnEn retrieves analogs using low-dimensional meteorological variables and forms an equally weighted ensemble. DF-AnEn uses the same deep features as the proposed method for retrieval but removes the physical wind-speed distance, temporal decay, and exponential weighting. Therefore, the comparison separates the effect of deep regime representation from that of multi-source weighted retrieval.

To keep the comparison focused, Supplementary Table S6 reports the 60-min horizon, where regime mismatch and uncertainty calibration are more challenging.

3.7 Cross-regional evaluation

To assess robustness and generalization across heterogeneous wind farms, Dataset B (Inner Mongolia) is introduced as an independent external test set. Training, validation, and hyper-parameter selection are conducted entirely on Dataset A (Zhejiang); Dataset B is reserved for external evaluation. Results are reported in Table 5 and Supplementary Tables S3–S5.

The cross-regional results show that the proposed method retains competitive deterministic accuracy under distribution shift from Zhejiang to Inner Mongolia. Ours achieves the lowest MAE at the 15- and 60-min horizons, whereas Persistence has lower MAE at 30 and 45 min. Ours achieves the lowest RMSE at 15 and 45 min. At 30 and 60 min, SVR has slightly lower RMSE (12.29 MW versus 12.62 MW, and 19.84 MW versus 20.14 MW, respectively). These results support competitive point accuracy alongside the probabilistic forecasts on the external dataset.

For probabilistic forecasting, Dataset B is clearly more challenging than Dataset A, as the empirical coverage of Ours is below the nominal levels: $PICP_{80}$ ranges from 73.80% to 77.65%, and $PICP_{90}$ ranges from 83.60% to 86.20%. This indicates under-coverage under cross-regional transfer and suggests that additional calibration would be needed for deployment on unseen wind farms. Nevertheless, Ours produces the sharpest intervals and the lowest Winkler scores at both 80% and 90% confidence levels across all horizons.

Table 5 Cross-regional forecasting performance at the 15- and 60-min horizons on Dataset B

Method	Horizon (min)	MAE	RMSE	R^2	$PICP_{80}$	$Winkler_{80}$	$PICP_{90}$	$Winkler_{90}$
Persistence	15	4.15	7.86	0.980	—	—	—	—
LSTM	15	6.70	11.27	0.960	—	—	—	—
Transformer	15	6.66	11.07	0.960	—	—	—	—
SVR	15	5.00	9.32	0.972	—	—	—	—
GRU	15	7.21	11.50	0.957	—	—	—	—
QR-LSTM	15	13.14	17.36	0.903	37.35	82.05	91.80	49.16
DeepAR	15	6.29	11.28	0.959	83.80	31.89	89.70	40.80
CKDE	15	7.18	11.96	0.954	78.20	39.17	84.40	49.53
Ours	15	3.85	7.02	0.984	73.80	20.71	86.20	26.99
Persistence	60	13.26	21.29	0.853	—	—	—	—
LSTM	60	13.83	20.91	0.859	—	—	—	—
Transformer	60	16.18	24.37	0.808	—	—	—	—
SVR	60	11.99	19.84	0.873	—	—	—	—
GRU	60	14.40	20.75	0.861	—	—	—	—
QR-LSTM	60	19.19	29.38	0.721	57.05	101.01	72.55	162.97
DeepAR	60	15.75	23.49	0.822	61.95	79.44	72.45	108.00
CKDE	60	13.71	20.40	0.866	76.15	66.11	82.65	88.53
Ours	60	11.38	20.14	0.869	75.40	59.73	83.60	84.17

4. Discussion

Compared with QR-LSTM and DeepAR, the proposed method does not improve probabilistic performance by simply widening the prediction intervals. Some deep probabilistic baselines achieve comparable short-horizon coverage, but their interval widths are substantially larger, which leads to higher Winkler scores.

Compared with CKDE, the main difference lies in how historical samples are conditioned and weighted. CKDE relies on low-dimensional conditioning variables and therefore struggles to match nonlinear operating regimes, especially at longer horizons and high confidence levels. By jointly using deep-feature distance, physical wind-speed proximity, and temporal continuity, the proposed method retrieves more relevant historical analogs and reconstructs sharper conditional densities within the non-parametric density-estimation framework.

The ramp-up case remains the most challenging scenario, with the proposed method achieving a $PICP_{80}$ of 64.49%, below the nominal 80% level. This lower coverage is mainly associated with the scarcity of upward-ramp samples, which account for only 5.4% of the training set, and with the steep power-curve region where small wind-speed deviations can lead to large power changes. Nevertheless, the proposed method still improves $PICP_{80}$ by 9.35, 13.09, and 14.96 percentage points over DeepAR, QR-LSTM, and CKDE, respectively, while producing much narrower intervals. Future work will explore ramp-event augmentation and adaptive calibration to improve coverage in this scenario.

The two components also play complementary roles in probabilistic quality. Multi-source weighting mainly sharpens the predictive intervals: at the 3-step horizon it reduces $PINAW_{80}$ from 0.162 to 0.126, trading a small amount of coverage for substantially sharper intervals. Physical-feature augmentation does not improve the probabilistic scores on its own. Consequently, the full model attains the best reliability–

sharpness trade-off, achieving the lowest Winkler_{80} at both horizons (48.23 at 3-step and 58.87 at 4-step) while keeping PICP_{80} close to the nominal 80% level.

Classical AnEn substantially underperforms the other two methods, indicating that hand-crafted low-dimensional distances are insufficient for identifying nonlinear wind-power regimes. The improvement from Classical AnEn to DF-AnEn shows the value of deep regime features. The comparison between DF-AnEn and the proposed method further shows that deep features alone are not sufficient: although DF-AnEn attains slightly higher PICP at the 90% level, it produces wider intervals. In contrast, the proposed method achieves lower PINAW and Winkler scores at both the 80% and 90% confidence levels, indicating a better reliability–sharpness trade-off. The proposed method therefore differs from DF-AnEn mainly in how the retrieved analogs are weighted and used for conditional density reconstruction.

This paper proposed a retrieval-augmented non-parametric framework for ultra-short-term probabilistic wind-power forecasting. The framework learns physically informed regime representations through a BiLSTM – Attention network, retrieves historical analogs using deep-feature similarity, wind-speed proximity, and temporal continuity, and reconstructs the conditional density through diffusion-smoothed weighted KDE. By doing so, it addresses the limited explicit use of historical regimes in end-to-end deep probabilistic models and the coarse similarity weighting of conventional non-parametric methods.

Experiments on the Zhejiang wind farm show that, at the 60-min horizon, the proposed method reduces MAE and RMSE by 21.2% and 22.7% relative to the Transformer baseline, and lowers the 80% Winkler score by 24.4% relative to CKDE. Under ramp and extreme-fluctuation regimes, it generally produces narrower intervals and lower Winkler scores than the compared probabilistic baselines, although upward-ramp coverage remains below the nominal level. The ablation and AnEn comparisons indicate that the two components should be interpreted jointly rather than as independent improvements. Cross-regional tests on the Inner Mongolia wind farm indicate that the framework retains competitive accuracy and favorable reliability–sharpness trade-offs under regional distribution shift, although additional calibration is still needed.

The neighbor-level analysis further shows that the proposed method provides traceable probabilistic forecasts by linking each prediction to physically relevant historical samples. However, two limitations remain. First, the empirical coverage under upward-ramp conditions is still below the nominal level, mainly because such events are scarce and difficult to calibrate. Future work will investigate ramp-event augmentation and adaptive interval calibration to improve coverage during rapid upward transitions. Second, the current framework retrieves analogs from a static historical pool; online updating and transfer-learning mechanisms will be explored to maintain performance under slowly drifting wind-farm conditions.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Wang, X., Guo, P. & Huang, X. A review of wind power forecasting models. *Energy Procedia* **12**, 770–778 (2011).

[2] Bošnjaković, M., Katinić, M., Santa, R. et al. Wind turbine technology trends. *Appl. Sci.* **12**, 8653 (2022).

[3] Korprasertsak, N. & Leephakpreeda, T. Robust short-term prediction of wind power generation under uncertainty via statistical interpretation of multiple forecasting models. *Energy* **180**, 387–397 (2019).

[4] Ma, Z. & Mei, G. A hybrid attention-based deep learning approach for wind power prediction. *Appl. Energy* **323**, 119608 (2022).

[5] Liu, X., Lin, Z. & Feng, Z. Short-term offshore wind speed forecast by seasonal ARIMA — a comparison against GRU and LSTM. *Energy* **227**, 120492 (2021).

[6] Wang, S., Zhang, N., Wu, L. et al. Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA–BP neural network method. *Renew. Energy* **94**, 629–636 (2016).

[7] Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D* **404**, 132306 (2020).

[8] Dey, R. & Salem, F. M. Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th MWSCAS, 2017, pp. 1597–1600.

[9] Zhao, Z., Yun, S., Jia, L. et al. Hybrid VMD–CNN–GRU-based model for short-term forecasting of wind power considering spatio-temporal features. *Eng. Appl. Artif. Intell.* **121**, 105982 (2023).

[10] Gong, M., Yan, C., Xu, W. et al. Short-term wind power forecasting model based on temporal convolutional network and Informer. *Energy* **283**, 129171 (2023).

[11] Meka, R., Alaeddini, A. & Bhaganagar, K. A robust deep learning framework for short-term wind power forecast of a full-scale wind farm using atmospheric variables. *Energy* **221**, 119759 (2021).

[12] Deng, Z., Zhang, X., Li, Z. et al. Probabilistic prediction of wind power based on improved Bayesian neural network. *Front. Energy Res.* **11**, 1309778 (2024).

[13] Chen, L., Yang, M., Guan, X. et al. Research on probabilistic wind power prediction based on QRCNN– BiLSTM model. In: 2025 4th ICEPET, 2025, pp. 468–472.

[14] Sun, Z., Shen, Y., Chen, Z. et al. Interval prediction method for wind speed based on ARQEA optimized by beta distribution and SWLSTM. *Front. Energy Res.* **10**, 927260 (2022).

[15] Rogers, T. J., Gardner, P., Dervilis, N. et al. Probabilistic modelling of wind turbine power curves with application of heteroscedastic Gaussian process regression. *Renew. Energy* **148**, 1124–1136 (2020).

[16] Capelletti, M., Raimondo, D. M. & De Nicolao, G. Wind power curve modeling: a probabilistic Beta regression approach. *Renew. Energy* **223**, 119970 (2024).

[17] Liu, L., Liu, J., Ye, Y. et al. Ultra-short-term wind power forecasting based on deep Bayesian model with uncertainty. *Renew. Energy* **205**, 598–607 (2023).

[18] Wen, H., Pinson, P., Ma, J. et al. Continuous and distribution-free probabilistic wind power forecasting: a conditional normalizing flow approach. *IEEE Trans. Sustain. Energy* **13**, 2250–2263 (2022).

[19] Zheng, Z., Yang, L. & Zhang, Z. Conditional variational autoencoder informed probabilistic wind power curve modeling. *IEEE Trans. Sustain. Energy* **14**, 2445–2460 (2023).

[20] Salinas, D., Flunkert, V., Gasthaus, J. et al. DeepAR: probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **36**, 1181–1191 (2020).

[21] Wang, Y., Xu, H., Zou, R. et al. A deep asymmetric Laplace neural network for deterministic and probabilistic wind power forecasting. *Renew. Energy* **196**, 497–517 (2022).

[22] Zhang, H., Yan, J., Liu, Y. et al. Multi-source and temporal attention network for probabilistic wind power prediction. *IEEE Trans. Sustain. Energy* **12**, 2205–2218 (2021).

[23] Yan, J., Li, K., Bai, E. et al. Special condition wind power forecasting based on Gaussian process and similar historical data. In: 2015 IEEE PES General Meeting, 2015, pp. 1–5.

[24] Wei, H., Chen, Y., Yu, M. et al. Alleviating distribution shift and mining hidden temporal variations for ultra-short-term wind power forecasting. *Energy* **290**, 130077 (2024).

[25] Gramacki, A. *Nonparametric Kernel Density Estimation and Its Computational Aspects* (Springer International Publishing, 2018).

[26] Jeon, J. & Taylor, J. W. Using conditional kernel density estimation for wind power density forecasting. *J. Am. Stat. Assoc.* **107**, 66–79 (2012).

[27] Wang, Z., Wang, W., Liu, C. et al. Short-term probabilistic forecasting for regional wind power using distance-weighted kernel density estimation. *IET Renew. Power Gener.* **12**, 1725–1732 (2018).

[28] Gu, B., Zhang, T., Meng, H. et al. Short-term forecasting and uncertainty analysis of wind power based on long short-term memory, cloud model and non-parametric kernel density estimation. *Renew. Energy* **164**, 687–708 (2021).

[29] Jeon, J., Panagiotelis, A. & Petropoulos, F. Probabilistic forecast reconciliation with applications to wind power and electric load. *Eur. J. Oper. Res.* **279**, 364–379 (2019).

[30] Wang, Y., Hu, Q., Srinivasan, D. et al. Wind power curve modeling and wind power forecasting with inconsistent data. *IEEE Trans. Sustain. Energy* **10**, 16–25 (2018).

[31] Delle Monache, L., Eckel, F. A., Rife, D. L. et al. Probabilistic weather prediction with an analog ensemble. *Mon. Weather Rev.* **141**, 3498–3516 (2013).

[32] Alessandrini, S., Delle Monache, L., Sperati, S. et al. A novel application of an analog ensemble for short-term wind power forecasting. *Renew. Energy* **76**, 768–781 (2015).

[33] Cervone, G., Clemente-Harding, L., Alessandrini, S. et al. Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. *Renew. Energy* **108**, 274–286 (2017).

[34] Junk, C., Delle Monache, L. & Alessandrini, S. Analog-based ensemble model output statistics. *Mon. Weather Rev.* **143**, 2909–2917 (2015).

[35] Yuan, D. D., Li, M., Li, H. Y. et al. Wind power prediction method: support vector regression optimized by improved jellyfish search algorithm. *Energies* **15**, 6404 (2022).

[36] Huang, S., Yan, C. & Qu, Y. Deep learning model—Transformer based wind power forecasting approach. *Front. Energy Res.* **10**, 1055683 (2023).

[37] Zhou, H., Zhang, S., Peng, J. et al. Informer: beyond efficient transformer for long sequence time-series forecasting. *In: Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 11106–11115 (2021).

[38] Peng, X., Wang, H., Lang, J. et al. EALSTM-QR: interval wind-power prediction model based on numerical weather prediction and deep learning. *Energy* **220**, 119692 (2021).

Acknowledgements

This work was supported in part by the Scientific and Technological Planning Project of Jilin Province (20250205077GH)

Author contributions

Hongbin Sun: Conceptualization, Methodology, Supervision, Funding acquisition, Writing – original draft. Shijun Guo: Software, Investigation, Visualization, Writing – original draft. Xin Pan: Validation, Formal analysis. Zhanpeng Qu: Writing – review & editing. Jianchuan Ma: Writing – review & editing. Yaqi Xu: Writing – review & editing.

Additional Information

Competing interests

The authors declare no competing interests.

Figure and table legends

Fig. 1. Overall framework of the proposed multi-source weighted probabilistic forecasting method.

Fig. 2. 80% prediction intervals during a ramp-intensive period (17 Jul 2022 09:00 – 18 Jul 2022 03:00).

Fig. 3. Retrieved-neighbor scatter and weighted predictive density for the ramp-up and ramp-down cases. The original panels a, c, d and f are shown.

Table 1. Key characteristics of the datasets.

Table 2. Point-forecast metrics at the 15- and 60-min horizons on Dataset A. MAE and RMSE are in MW. Complete results for all four horizons are in Supplementary Table S7.

Table 3. Multi-horizon probabilistic metrics on Dataset A at the 80% and 90% nominal levels. PICP is in %, PINAW is normalized and Winkler is in MW.

Table 4. Component-ablation metrics at the 60-min horizon on Dataset A. MAE, RMSE and Winkler are in MW; PICP is in %, and PINAW is normalized.

Table 5. Selected 15- and 60-min cross-regional results on Dataset B. MAE, RMSE and Winkler are in MW; PICP is in %. Full horizons and PINAW values are retained in Supplementary Tables S3–S5. A dash indicates an unreported metric.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SR.docx