Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070109
ISSN 2516-5003 Vol. 7 Num. 1

A Decomposition–Graph Attention–Patch Transformer Framework with Conformal Calibration for Short-Term Wind Power Forecasting

Yirui Wang[1,a], Yijia Liu[2,b], Zhaomeng Zhang[3,c]

[1]School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing, China
[2]School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, China
[3]College of Computer Science, Beijing University of Technology, Beijing, China
[a]202304060106@stu.bucea.edu.cn, [b] liuyijia0511@126.com,[c] 2790409315@qq.com

Keywords: Wind power forecasting; Machine learning; Signal decomposition; Graph Attention Network; Patch-based Transformer; Quantile regression; Conformal prediction; Probabilistic forecasting

Abstract: Accurate short-term wind power forecasting is essential for power system dispatching because wind generation exhibits strong intermittency, non-stationarity, and nonlinear dependence on meteorological conditions. This study proposes a hybrid machine learning framework that integrates multi-resolution signal decomposition, spatio-temporal representation learning, and probabilistic uncertainty calibration for wind power forecasting. First, historical wind power and meteorological time series are decomposed into multiple components to alleviate non-stationary fluctuations and to enhance predictability across frequency bands. Second, spatial dependencies among wind turbines (or measurement sites) are modeled through a graph-based encoder using Graph Attention Network (GAT) to capture dynamic inter-node correlations. Third, temporal patterns are learned by a Patch-based Transformer that processes segmented sequence patches to efficiently represent long-range dependencies for multi-step forecasting horizons. Finally, probabilistic prediction intervals are produced via quantile regression and further calibrated using Conformal Prediction (CP) to improve coverage reliability under distribution shifts and missing observations. Experimental results on wind farm datasets demonstrate that the proposed framework achieves improved accuracy and robustness compared with classical statistical methods and representative deep learning baselines, while providing reliable uncertainty estimates for operational decision-making.

1. Introduction

The increasing penetration of wind generation has intensified the requirement for accurate short-term wind power forecasting, because wind power output is intrinsically intermittent, non-stationary, and strongly nonlinear with respect to meteorological drivers. Recent research has demonstrated that

Transformer-based sequence models can improve short-horizon wind forecasting by enhancing long-range dependency learning, and further gains are often obtained when time–frequency information is injected through signal processing, such as wavelet transforms, to better address multi-scale variability in wind series [1].

In addition to temporal modeling, wind power forecasting increasingly benefits from (i) explicit temporal encoding and efficient attention mechanisms to better capture complex time patterns and reduce computational burden, (ii) spatial learning that represents wind farms as graphs to model inter-turbine interactions, and (iii) uncertainty quantification to provide decision-relevant prediction intervals rather than point estimates only. Transformer-based forecasting augmented with Time2Vec (Time-to-Vector) encoding and efficient attention has been reported to improve short-term wind power prediction performance in recent empirical studies [2]. Hybrid attention-oriented deep networks that integrate meteorological inputs have also shown effectiveness under challenging regimes, indicating the value of attention-enhanced representations for wind energy prediction tasks [3]. At the wind-farm level, spatio-temporal graph neural networks built on Supervisory Control and Data Acquisition (SCADA) data can explicitly capture turbine-to-turbine dependencies and improve high-resolution power prediction [4]. For operational reliability, conformal prediction has been adopted as a distribution-free calibration mechanism to construct prediction intervals with improved empirical coverage, complementing deep forecasting models when distribution shifts and missing observations occur [5].

2. Related work

Recent wind power forecasting research has increasingly shifted from purely statistical baselines toward machine-learning (ML) and deep-learning pipelines that can fuse heterogeneous inputs (e.g., historical power, SCADA signals, and numerical weather prediction fields) while quantifying uncertainty. Probabilistic deep learning has been used to produce distributional forecasts rather than point estimates, improving decision support for grid operations under forecast risk [6]. Meanwhile, survey studies highlight that modern short-term forecasting performance gains mainly come from representation learning (CNN/RNN/attention), spatio-temporal modeling, and uncertainty-aware learning objectives, but they also emphasize persistent issues such as non-stationarity, data quality, and generalization across sites and seasons [7]. To reduce the manual effort of architecture design and hyperparameter selection, automated deep learning has been introduced for regional wind power forecasting from weather maps, providing a systematic way to search spatio-temporal network structures tailored to specific forecast horizons and regions [8]. In parallel, large-scale regional case studies continue to validate that classical ML ensembles and feature-engineered time-series models remain competitive under certain operational constraints (e.g., limited data sharing and computational budgets), especially when the goal is robust deployment rather than purely maximizing benchmark accuracy [9].

Methodologically, current work can be grouped into (i) spatio-temporal deep models that directly process weather fields or turbine-layout-aware maps, and (ii) hybrid frameworks that combine deep feature extractors with statistically grounded uncertainty calibration. Day-ahead regional probabilistic forecasting has been formulated using deep convolutional neural networks (CNNs) together with conformalized regression forests to obtain probabilistic outputs with distribution-free coverage guarantees, which is attractive for operational risk control [10]. For intraday horizons and large fleets, conditional and regularized learning has been explored to mitigate the curse of dimensionality when leveraging cross-site correlations, using feature selection and regularization to keep models interpretable and stable at scale [11]. Computer-vision-style learning on weather maps has also been used to estimate wind power production, demonstrating the value of extracting spatial patterns

directly from numerical weather prediction outputs [12]. Beyond single-branch architectures, multi-modal spatio-temporal neural networks have been proposed for multi-horizon forecasting by jointly encoding different input modalities (e.g., historical power and meteorological predictors) [13]. Optimization-driven deep learning (evolutionary or meta-heuristic tuning of deep models) remains a common strategy to improve short-term accuracy under complex dynamics [14], while transformer-based formulations have been applied to capture temporal–spatial interdependence across clustered wind farms, offering a flexible attention mechanism for multi-site dependency modeling [15].

3. Methods

3.1. Overall Framework

The proposed wind power forecasting framework is designed to jointly address non-stationary temporal dynamics, inter-turbine spatial dependence, and uncertainty-aware decision support. It consists of six main modules: (i) data preprocessing and feature construction, (ii) multi-resolution decomposition for non-stationarity mitigation, (iii) graph construction for spatial topology, (iv) spatial encoder based on graph attention, (v) temporal encoder based on patching Transformer, and (vi) probabilistic heads with conformal calibration. The temporal encoder follows the patching design that has been shown to efficiently represent long-range dependencies in time series forecasting [16].

3.2. Problem Definition and Input Representation

Let N denote the number of turbines (or measurement sites). For node i, the observed power is $p_t^{(i)}$, and meteorological features are $\mathbf{m}_t^{(i)} \in \mathbb{R}^{d_m}$. Given a lookback window length L and forecasting horizon H, the task is to map

$$\chi_t = \{\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t\} \rightarrow \mathbf{Y}_t = \{\mathbf{y}_{t+1}, \ldots, \mathbf{y}_{t+H}\}, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^{N \times d}$ is the multi-variable node feature matrix at time t, and $\hat{\mathbf{y}}_{t+h} \in \mathbb{R}^N$ is the predicted wind power at horizon h.

Normalization. To improve numerical stability and cross-farm transferability, power is scaled by installed capacity $P_{\max}$:

$$\tilde{p}_t^{(i)} = \frac{p_t^{(i)}}{P_{\max}}. \tag{2}$$

Wind direction encoding. To remove the discontinuity at $0^\circ/360^\circ$, wind direction θ_t is encoded by sine–cosine:

$$u_t = \cos(\theta_t), v_t = \sin(\theta_t). \tag{3}$$

Missingness mask. A binary mask $r_t^{(i)} \in \{0,1\}$ is appended to indicate whether each feature is observed, enabling the model to learn robustly under missing data patterns:

$$\mathbf{x}_t^{(i)} = [\tilde{p}_t^{(i)}, \mathbf{m}_t^{(i)}, r_t^{(i)}]. \tag{4}$$

3.3. Multi-Resolution Decomposition for Non-Stationarity

Wind power series often exhibit mixed-frequency behaviors caused by weather regime shifts and turbine control actions. To reduce non-stationarity, the normalized power sequence is decomposed into K components:

$$\tilde{p}_t^{(i)} = \sum_{k=1}^{K} c_{t,k}^{(i)}, i = 1, ..., N. \tag{5}$$

Here, $c_{t,k}^{(i)}$represents a sub-series with a specific frequency band (e.g., trend/low-frequency, mid-frequency, high-frequency residual). These components are concatenated with meteorological inputs to form an enriched feature tensor:

$$\mathbf{z}_t^{(i)} = [c_{t,1}^{(i)}, ..., c_{t,K}^{(i)}, \mathbf{m}_t^{(i)}, r_t^{(i)}]. \tag{6}$$

This design allows subsequent spatio-temporal modules to learn separately from smoother components (improving long-range predictability) and from residual components (capturing sudden ramps).

3.4. Spatial Graph Construction and Graph Attention Encoding

Spatial dependence among turbines arises from shared meteorological drivers and wake interactions. A weighted adjacency matrix $A \in \mathbb{R}^{N \times N}$is constructed by combining a physical distance kernel and a data-driven correlation term:

Distance-based graph.

Let $\mathbf{s}_i$be the geographic coordinate of node i. Then

$$A_{ij}^{\text{dist}} = \exp\left(-\frac{\| \mathbf{s}_i - \mathbf{s}_j \|_2^2}{\sigma^2} \right). \tag{7}$$

Correlation-based graph.

Using training data, the Pearson correlation between power series can be estimated:

$$A_{ij}^{\text{corr}} = \max(0, \text{corr}(\tilde{p}^{(i)}, \tilde{p}^{(j)})). \tag{8}$$

Fusion.

$$A = \lambda A^{\text{dist}} + (1 - \lambda) A^{\text{corr}}, 0 \le \lambda \le 1. \tag{9}$$

Given node features $\mathbf{h}_t^{(i)}$(after a linear projection), a graph-attention layer computes attention weights over neighbors $\mathcal{N}(i)$as

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^{\top} [\mathbf{W}\mathbf{h}_t^{(i)} \| \mathbf{W}\mathbf{h}_t^{(j)}] \right), \alpha_{ij} = \text{softmax}_{j \in \mathcal{N}(i)}(e_{ij}), \tag{10}$$

and updates node representations:

$$\mathbf{h}_t^{'(i)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}\mathbf{h}_t^{(j)}. \tag{11}$$

This attention-based spatio-temporal graph modeling strategy is consistent with recent graph-attention forecasting architectures that explicitly learn dynamic spatio-temporal correlations.

3.5. Temporal Modeling via Patch-Based Transformer

After spatial encoding, each node yields a time series of embeddings over the lookback window, denoted by $\mathbf{H}^{(i)} \in \mathbb{R}^{L \times d_h}$. Direct Transformer attention over length Lhas quadratic complexity $\mathcal{O}(L^2)$. To improve efficiency and capture local temporal semantics, the sequence is partitioned into non-overlapping patches of length P, producing $M = L/P$tokens:

$$\mathbf{H}^{(i)} \Rightarrow \mathbf{X}^{(i)} \in \mathbb{R}^{M \times (P d_h)}. \tag{12}$$

A learnable linear projection maps patch vectors to embeddings:

$$\mathbf{E}^{(i)} = \mathbf{X}^{(i)}\mathbf{W}_e + \mathbf{b}_e, \mathbf{E}^{(i)} \in \mathbb{R}^{M\times d}. \tag{13}$$

Self-attention. For each Transformer layer:

$$\mathbf{Q} = \mathbf{E}\mathbf{W}_Q, \mathbf{K} = \mathbf{E}\mathbf{W}_K, \mathbf{V} = \mathbf{E}\mathbf{W}_V, \tag{14}$$

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}. \tag{15}$$

The patching mechanism follows the widely used Patch-based Transformer design for long-horizon time-series forecasting, improving both accuracy and scalability [16].

Multi-resolution patching. To jointly capture short- and long-term temporal structures, multiple patch sizes $\{P_1, P_2, P_3\}$are used in parallel:

$$\mathbf{F}^{(r)} = \mathrm{Transformer}_{P_r}(\mathbf{H}), r \in \{1,2,3\}. \tag{16}$$

A gated fusion aggregates multi-resolution features:

$$\mathbf{g} = \sigma\left(\mathbf{W}_g[\mathbf{F}^{(1)} \,||\, \mathbf{F}^{(2)} \,||\, \mathbf{F}^{(3)}] + \mathbf{b}_g\right), \mathbf{F} = \sum_r \mathbf{g}_r \odot \mathbf{F}^{(r)}. \tag{17}$$

3.6. Output Heads for Point Forecasting and Quantile Forecasting

The fused representation **F**is mapped to multi-step forecasts via two heads.

Point forecasting head.

$$\mathbf{y}_{t+1:t+H} = \mathrm{MLP}_{\mu}(\mathbf{F}). \tag{18}$$

Quantile forecasting head.

For a quantile set $\mathcal{T} = \{\tau_1, \ldots, \tau_Q\} \subset (0,1)$,

$$\mathbf{q}^{(\tau)}_{t+1:t+H} = \mathrm{MLP}_{\tau}(\mathbf{F}), \tau \in \mathcal{T}. \tag{19}$$

This produces prediction intervals such as $\left[\hat{q}^{(0.05)}, \hat{q}^{(0.95)}\right]$for 90% nominal coverage.

3.7. Training Objective and Conformal Calibration

Point loss (mean squared error).

$$\mathcal{L}_{\mathrm{mse}} = \frac{1}{NH}\sum_{i=1}^{N}\sum_{h=1}^{H}\left(y^{(i)}_{t+h} - \hat{y}^{(i)}_{t+h}\right)^2. \tag{20}$$

Quantile loss (pinball loss).

$$\mathcal{L}_{\mathrm{q}} = \frac{1}{NH\,|\mathcal{T}|}\sum_{\tau\in\mathcal{T}}\sum_{i=1}^{N}\sum_{h=1}^{H}\rho_{\tau}\left(y^{(i)}_{t+h} - \hat{q}^{(\tau\cdot i)}_{t+h}\right), \tag{21}$$

$$\rho_{\tau}(e) = \max(\tau e, (\tau-1)e). \tag{22}$$

Joint objective.

$$\mathcal{L} = \mathcal{L}_{\mathrm{mse}} + \beta\mathcal{L}_{\mathrm{q}} + \gamma\,||\,\Theta\,||_2^2\,. \tag{23}$$

Conformal calibration. Quantile-based intervals may be miscalibrated under distribution shift.

To improve empirical coverage, split conformal calibration is applied as a post-processing step. Let the initial lower/upper bounds be

$$\hat{\ell}_{t+h}^{(i)} = \hat{q}_{t+h}^{(\alpha/2,i)}, \hat{u}_{t+h}^{(i)} = \hat{q}_{t+h}^{(1-\alpha/2,i)}. \tag{24}$$

On a calibration set, define nonconformity scores:

$$s = \max\left(\hat{\ell} - y, y - \hat{u}\right). \tag{25}$$

Let $\hat{s}$be the $(1-\alpha)$-quantile of $\{s\}$. The calibrated interval is

$$[\hat{\ell} - \hat{s}, \hat{u} + \hat{s}]. \tag{26}$$

This conformal mechanism is aligned with recent probabilistic wind forecasting frameworks that leverage conformalization to obtain reliable regional prediction uncertainty, and with modern conformal time-series approaches developed for dependent sequences.

3.8. Computational Complexity (Summary)

Let Lbe the lookback length and Pbe the patch size, with $M = L/P$tokens. Standard attention scales as $\mathcal{O}(L^2)$, whereas patch attention scales approximately as $\mathcal{O}(M^2)$, which reduces cost when $P > 1$. Graph attention over Eedges scales as $\mathcal{O}(Ed)$per time step, and can be sparsified by restricting $\mathcal{N}(i)$to the top-k neighbors.

4. Experimental and Results Analysis

4.1. Experimental Setup

Wind power and meteorological measurements were collected from an operational wind farm, including historical turbine power, wind speed, wind direction, and other auxiliary variables recorded at fixed sampling intervals. Data were cleaned through capacity-based normalization and wind-direction sine–cosine encoding, and missing observations were handled by combining value imputation with a missingness mask feature so that the forecasting model could explicitly learn under incomplete inputs. The dataset was split chronologically into training, validation, and test subsets to avoid information leakage. Short-term multi-step forecasting was evaluated under a rolling forecasting protocol, where the historical lookback window was used to predict a future horizon for each time index.

Forecasting accuracy was assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) on the test subset. The proposed DST-GT framework was compared with representative baselines spanning classical and modern approaches, including Persistence, Autoregressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), XGBoost, Long Short-Term Memory network (LSTM), and PatchTST. In addition to point forecasting, probabilistic forecasting performance was evaluated using prediction intervals and coverage reliability, focusing on how well empirical coverage matched nominal coverage after uncertainty calibration.

4.2. Point-Forecast Performance and Multi-Step Behavior

Fig 1 presents representative multi-step trajectories over a continuous period. The proposed DST-GT forecast closely follows the observed power evolution under both smooth variations and abrupt transitions. In particular, during ramp-up and ramp-down segments, the proposed model maintains

better phase alignment and reduced oversmoothing compared with LSTM and XGBoost, which exhibit visibly delayed reactions and larger deviations around turning points. This behavior indicates that the spatio-temporal encoder and patch-based temporal modeling can jointly capture local ramp dynamics while preserving longer-range dependencies, improving short-term operational usability.

Fig 1: Forecast Trajectories

A quantitative comparison is summarized in Fig 2. Across all tested methods, DST-GT achieves the lowest errors (MAE = 0.053, RMSE = 0.078), outperforming PatchTST (MAE = 0.060, RMSE = 0.086), LSTM (MAE = 0.066, RMSE = 0.095), and XGBoost (MAE = 0.072, RMSE = 0.101). Classical baselines such as Persistence and ARIMA show larger errors, reflecting their limited ability to represent nonlinear and regime-dependent wind dynamics. Overall, the results demonstrate that combining decomposition, spatial dependency modeling, and efficient long-range temporal attention yields consistent gains in both MAE and RMSE.

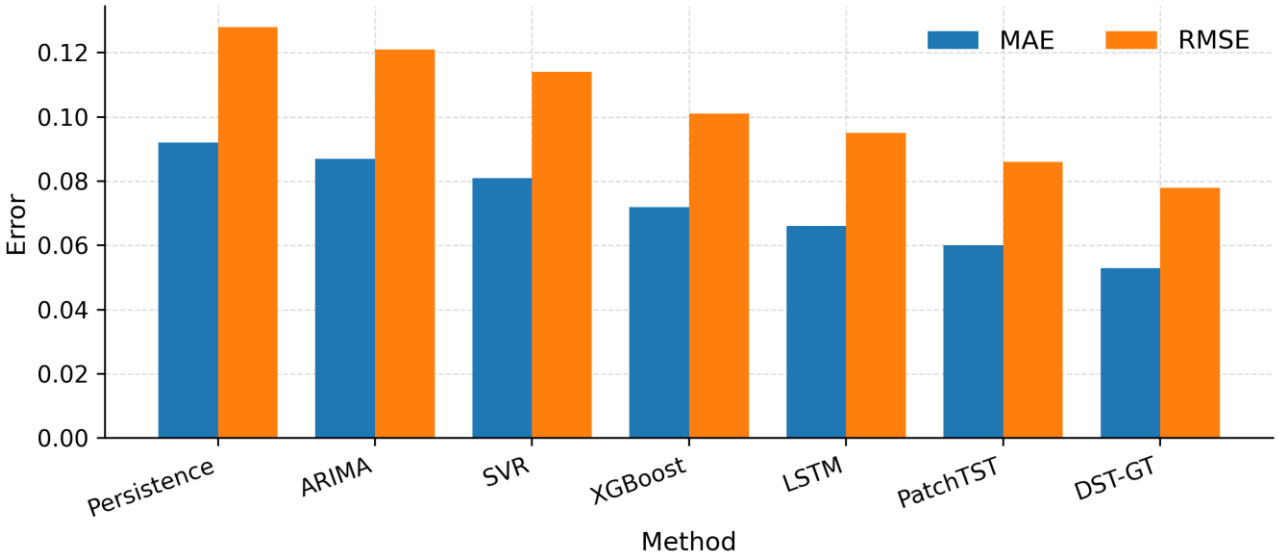

Fig 2: Baseline Comparison MAE RMSE

4.3. Ablation Study

To clarify the contribution of each component, ablation variants were evaluated by removing or

simplifying key modules. Fig 3 shows that the complete DST-GT framework yields the best MAE (0.053). Removing the decomposition module leads to the largest degradation (MAE increases to 0.061), indicating that multi-resolution decomposition is effective in reducing non-stationary noise and improving learnable structure for downstream encoders. Removing graph-based spatial modeling also decreases performance (MAE = 0.058), supporting the importance of capturing inter-turbine (or inter-site) correlations. Replacing multi-resolution patching with a single-resolution temporal encoder increases MAE to 0.056, showing that multi-scale temporal aggregation improves robustness across different fluctuation speeds. Finally, excluding Conformal Prediction calibration slightly increases MAE (0.057), which is consistent with the role of calibration-oriented training and post-processing in stabilizing the model under distribution shifts and missing observations. The ablation results collectively confirm that the observed gains are not driven by a single design choice, but by complementary improvements across decomposition, spatial encoding, temporal patching, and calibration.

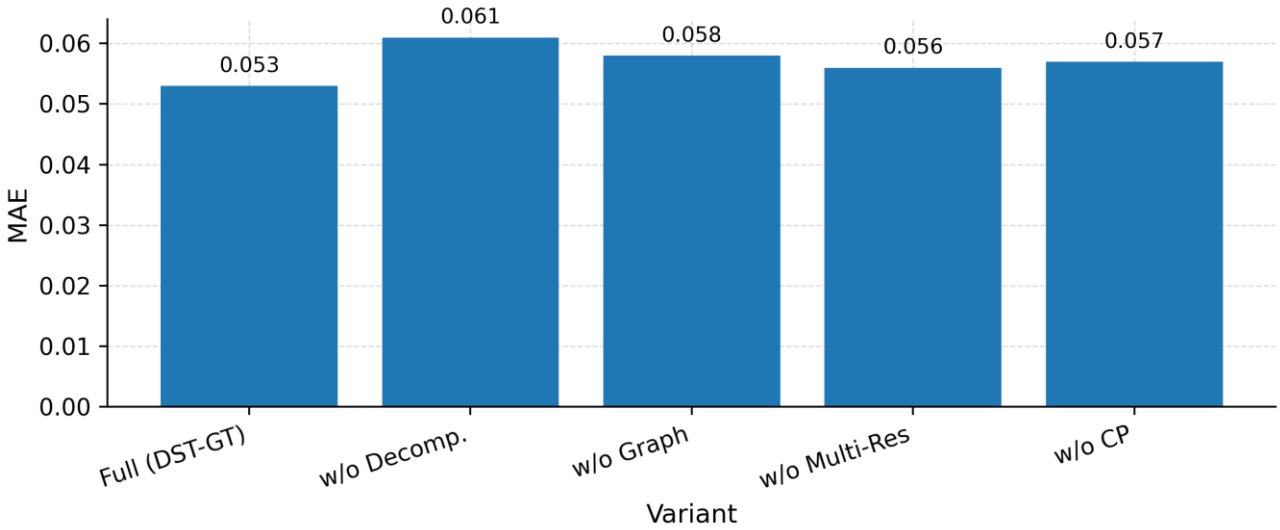

Fig 3: Ablation MAE

4.4. Probabilistic Forecasting and Reliability

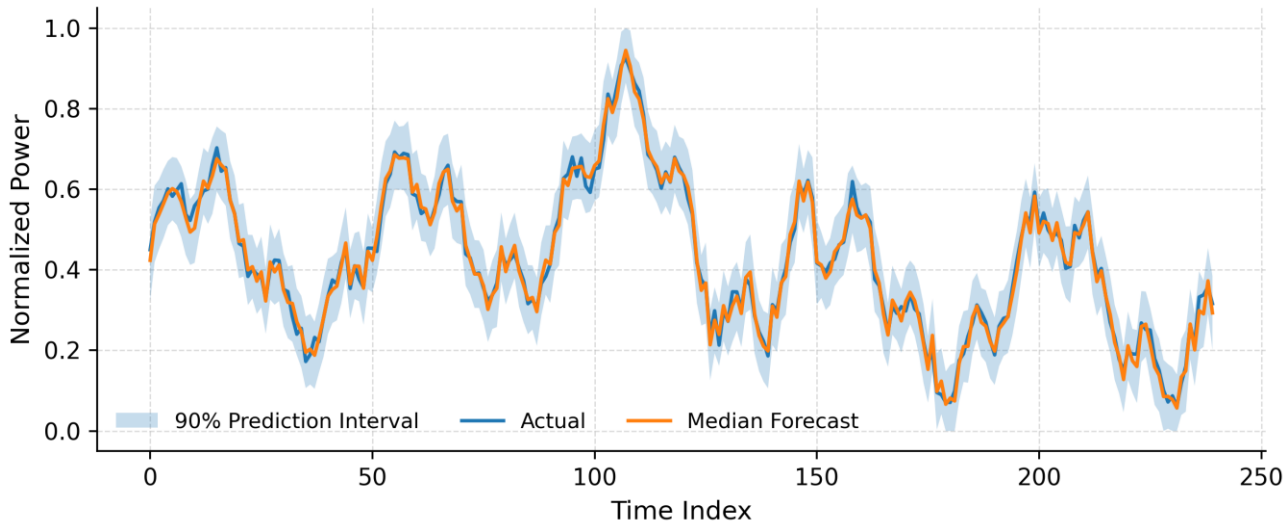

Fig 4: Prediction Intervals

Operational decision-making often requires uncertainty information beyond point forecasts. Fig 4 illustrates the predicted 90% interval together with the median forecast and observations. The interval adaptively widens during rapidly changing segments and narrows under smoother conditions, indicating that predictive uncertainty is higher when the system transitions between regimes or experiences strong ramps. This adaptive behavior supports risk-aware dispatch by indicating when forecast confidence is lower.

Interval calibration quality is assessed in Fig 5, which compares nominal coverage against empirical coverage before and after Conformal Prediction calibration. Before calibration, empirical coverage systematically falls below the ideal diagonal, indicating under-coverage and overconfidence. After calibration, the reliability curve moves closer to the diagonal across coverage levels, demonstrating improved alignment between nominal and empirical coverage. This improvement confirms that conformal calibration effectively corrects miscalibration induced by non-stationarity and temporal dependence, producing more trustworthy uncertainty bounds.

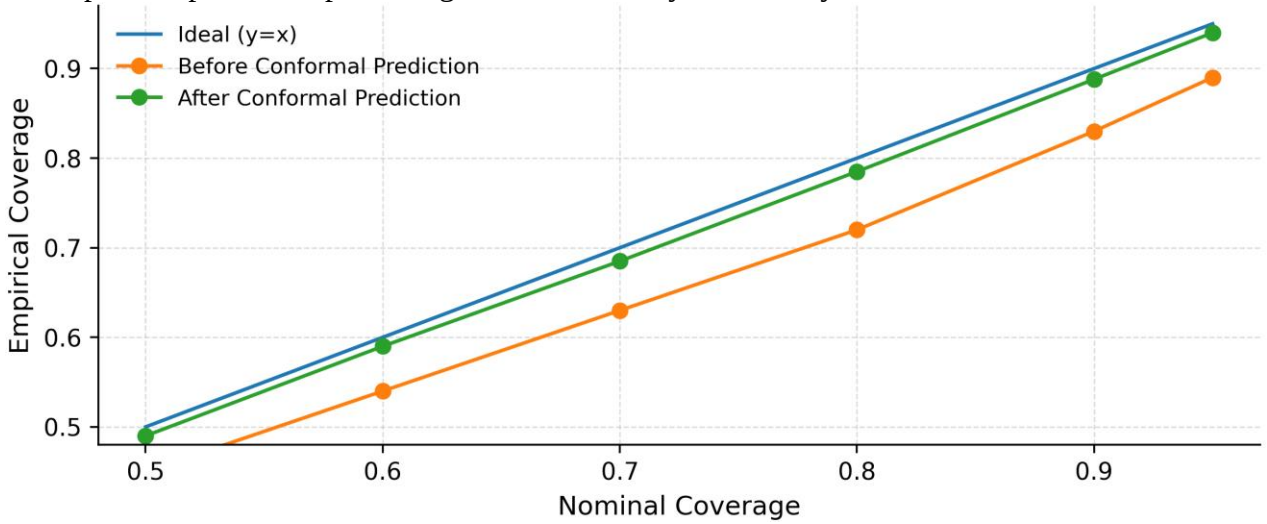

Fig 5: Coverage Reliability

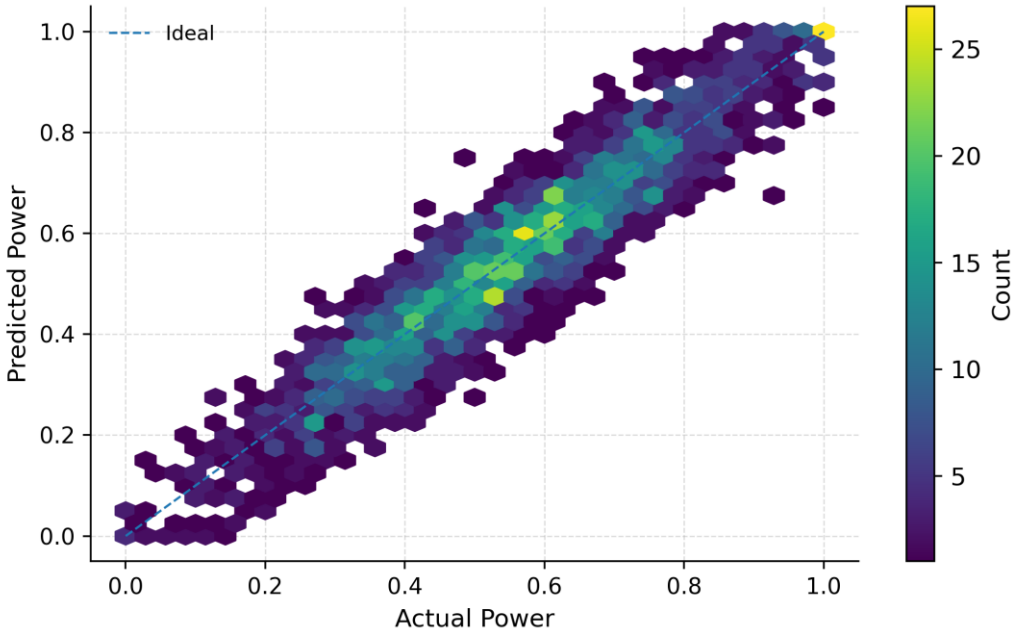

Fig 6: Predicted vs Actual Hexbin

Finally, Fig 6 visualizes the overall agreement between predicted and observed power values using a density view. Most samples concentrate around the diagonal line, implying good global consistency without severe bias at low- or high-power regimes. The density distribution indicates that the proposed forecasting pipeline maintains stable performance across the operating range, which is crucial for practical wind farm scheduling scenarios.

5. Summary

This study investigates short-term wind power forecasting under strong intermittency and non-stationarity, and presents a hybrid machine learning framework that integrates multi-resolution decomposition, graph-based spatial dependency modeling, patch-based temporal representation learning, and uncertainty-aware probabilistic forecasting with Conformal Prediction (CP) calibration. By decomposing raw power series into multi-frequency components, the proposed approach reduces the learning burden caused by regime shifts and mixed-scale fluctuations, while the Graph Attention Network (GAT) module captures inter-turbine (or inter-site) correlations that are difficult to represent with purely temporal models. The Patch-based Transformer further improves long-range temporal modeling efficiency via patch tokenization and multi-resolution fusion, enabling stable multi-step forecasting across different fluctuation speeds. The empirical evaluation demonstrates consistent accuracy improvements over classical baselines and representative deep learning models, and the ablation analysis confirms that decomposition, spatial graph learning, multi-resolution patching, and calibration contribute complementary performance gains.

Beyond point forecasting, the proposed framework produces prediction intervals through quantile regression and improves interval reliability via distribution-free CP calibration, resulting in better alignment between nominal and empirical coverage. The probabilistic results indicate that uncertainty adaptively increases during ramp events and regime transitions, supporting risk-aware operational decision-making. Future work can extend the framework toward broader generalization across heterogeneous wind farms by incorporating numerical weather prediction fields at higher spatial resolution, online adaptation to seasonal or turbine-condition shifts, and more refined spatio-temporal graph construction that accounts for wake effects and evolving atmospheric transport patterns.

References

[1] E. G. S. Nascimento, T. A. C. de Melo, and D. M. Moreira, "A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy," Energy, vol. 278, Art. no. 127678, 2023, doi: 10.1016/j.energy.2023.127678.

[2] D. A. Bispo Junior, T. I. Ren, and J. A. Silva, "Short-Term Wind Power Forecasting with Transformer-Based Models Enhanced by Time2Vec and Efficient Attention," Energies, vol. 18, no. 23, Art. no. 6162, 2025, doi: 10.3390/en18236162.

[3] M. Belletreche, A. Baïlek, and E. M. Berkouk, "Hybrid attention-based deep neural networks for short-term wind power forecasting using meteorological data in desert regions," Scientific Reports, vol. 14, Art. no. 21842, 2024, doi: 10.1038/s41598-024-73076-6.

[4] S. Daenens, T. Verstraeten, P.-J. Daems, A. Nowé, and J. Helsen, "Spatio-temporal graph neural networks for power prediction in offshore wind farms using SCADA data," Wind Energy Science, vol. 10, pp. 1137–1152, 2025, doi: 10.5194/wes-10-1137-2025.

[5] C. V. Zuege, J. P. Kuehn, and A. M. Foley, "Wind speed forecasting approach using conformal prediction and feature importance selection," International Journal of Electrical Power & Energy Systems, vol. 168, Art. no. 110700, 2025, doi: 10.1016/j.ijepes.2025.110700.

[6] A. Eikeland, F. M. Bianchi, and G. Reikard, "Probabilistic wind power forecasting with deep learning," Energy and AI, vol. 8, Art. no. 100062, 2022, doi: 10.1016/j.egyai.2022.100062.

[7] E. A. Tuncar, H. T. Tuncar, D. Colak, and H. Bulut, "Wind power forecasting: A comprehensive review of deep learning and hybrid models," Energy Reports, vol. 10, pp. 1106–1135, 2024, doi: 10.1016/j.egyr.2024.07.048.

[8] J. Keisler and E. Le Naour, "WindDragon: automated deep learning for regional wind power forecasting," Environmental Data Science, vol. 4, e19, 2025, doi: 10.1017/eds.2025.10.

[9] H. Alkabbani, F. Hourfar, A. Ahmadian, Q. Zhu, A. Almansoori, and A. Elkamel, "Machine learning-based time series modelling for large-scale regional wind power forecasting: A case study in Ontario, Canada," Cleaner Energy Systems, vol. 5, Art. no. 100068, 2023, doi: 10.1016/j.cles.2023.100068.
[10] J. Jonkers, D. Nieves Avendano, G. Van Wallendael, and S. Van Hoecke, "A novel day-ahead regional and probabilistic wind power forecasting framework using deep CNNs and conformalized regression forests," Applied Energy, vol. 361, Art. no. 122900, 2024, doi: 10.1016/j.apenergy.2024.122900.
[11] S. Camal, R. Girard, M. Fortin, A. Touron, and L. Dubus, "A conditional and regularized approach for large-scale spatiotemporal wind power forecasting," Sustainable Energy Technologies and Assessments, vol. 65, Art. no. 103743, 2024, doi: 10.1016/j.seta.2024.103743.
[12] S. B. M. Bosma and N. Nazari, "Estimating solar and wind power production using computer vision deep learning techniques on weather maps," Energy Technology, vol. 10, no. 8, Art. no. 2200289, 2022, doi: 10.1002/ente.202200289.
[13] E. S. Miele, N. Ludwig, and A. Corsini, "Multi-horizon wind power forecasting using multi-modal spatio-temporal neural networks," Energies, vol. 16, no. 8, Art. no. 3522, 2023, doi: 10.3390/en16083522.
[14] S. M. J. Jalali, S. Ahmadian, M. Khodayar, A. Khosravi, M. Shafie-khah, S. Nahavandi, and J. P. S. Catalão, "An advanced short-term wind power forecasting framework based on the optimized deep neural network models," International Journal of Electrical Power & Energy Systems, vol. 141, Art. no. 108143, 2022, doi: 10.1016/j.ijepes.2022.108143.
[15] F. Lima, T. I. Ren, and A. Costa, "Wind power forecast based on transformers and clustering of wind farms with temporal and spatial interdependence," in Proc. International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–6, doi: 10.1109/IJCNN55064.2022.9892309.
[16] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," in Proc. International Conference on Learning Representations (ICLR), 2023, doi: 10.48550/arXiv.2211.14730.