GSTAformer: graph-guided spatio-temporal autoformer for mid-term wind power forecasting

Yuan, S. ORCID: https://orcid.org/0009-0008-2792-9772, Mao, Y. ORCID: https://orcid.org/0009-0007-9967-6770, Tian, C., Yu, F., Guo, T. and Xia, M. ORCID: https://orcid.org/0000-0003-4681-9129 (2026) GSTAformer: graph-guided spatio-temporal autoformer for mid-term wind power forecasting. Energies, 19 (1). 254. ISSN 1996-1073 doi: 10.3390/en19010254 Available at https://centaur.reading.ac.uk/127890/

energies

MDPI

Article

GSTAformer: Graph-Guided Spatio-Temporal Autoformer for Mid-Term Wind Power Forecasting

Shi Yuan [1], Yulu Mao [1], Chenyu Tian [1], Fang Yu [2], Tengyue Guo [1,3] and Min Xia [1,*]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, No. 219, Ningliu Road, Nanjing 210044, China; 202312490003@nuist.edu.cn (S.Y.); 202312490010@nuist.edu.cn (Y.M.); 202312490631@nuist.edu.cn (C.T.); wj835167@student.reading.ac.uk (T.G.)
[2] China Electric Power Research Institute Co., Ltd., No. 8 Nanrui Road, Gulou District, Nanjing 210000, China; yufang@epri.sgcc.com.cn
[3] Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK
* Correspondence: xiamin@nuist.edu.cn

Abstract

Accurate wind power forecasting is crucial for modern power systems, yet most deep learning models neglect spatial relationships between turbines. We propose GSTAformer, a graph-guided spatio-temporal model capturing both spatial and temporal dependencies through MIC- and PCC-built graphs; GraphSAGE for spatial feature extraction; multi-scale convolution for trend detection; and an improved Autoformer for temporal modeling. Experiments on SDWPF and GEFCom2012 datasets demonstrate GSTAformer's superior performance, achieving a 24 h mean squared error (MSE) of 0.7480 and mean absolute error (MAE) of 0.6362 on SDWPF. This work integrates graph-based spatial modeling with enhanced temporal forecasting for medium-term wind power prediction, providing a coherent framework suited to complex wind energy scenarios.

Keywords: wind power forecasting; spatio-temporal modeling; graph neural network; Autoformer; GraphSAGE; multi-scale convolution; medium-term prediction

1. Introduction

Amid global fossil fuel depletion and growing environmental concerns, renewable energy development has become a global priority. As a clean and sustainable resource, wind power has received wide attention for its availability and near-zero operational emissions [1]. However, its stochastic and intermittent nature—strongly influenced by meteorological and geographical factors—introduces significant uncertainty to power system operation, creating challenges for grid stability and energy scheduling [2].

For practical power system management, accurate medium-term wind power forecasting (ranging from several hours to a few days ahead) is particularly important. It supports day-ahead scheduling, power trading, and maintenance planning, helping operators improve system reliability and economic efficiency [3]. Accurate 6–24 h wind forecasts are also essential in practical decision frameworks such as day-ahead bidding and coordinated electricity–carbon market operation, where recent studies show that market efficiency is highly dependent on renewable forecasting accuracy [4]. However, medium-term forecasting is more challenging than short-term forecasting due to higher uncertainty and error accumulation over longer horizons. Therefore, developing accurate and robust

Academic Editors: Kangji Li and Wenping Xue

Received: 2 December 2025
Revised: 18 December 2025
Accepted: 30 December 2025
Published: 2 January 2026

medium-term forecasting methods is of great significance for wind power integration and system operation.

Over the past decade, diverse forecasting approaches have been developed, ranging from physical and statistical models to machine learning and deep learning frameworks [5,6]. Deep learning has shown superior ability to capture nonlinear and temporal dependencies in wind data. Early works mainly used recurrent or probabilistic networks such as long short-term memory (LSTM), gated recurrent unit (GRU), and deep belief network (DBN). For instance, Xiong et al. [7] combined convolutional neural networks (CNNs) and attention with LSTM, and Boucetta et al. [8] employed GRU with attention for short-term prediction. These studies laid a foundation for data-driven forecasting but still fail to represent spatial dependencies across turbines.

Recent efforts emphasize joint spatio-temporal modeling, motivated by the correlation between turbine locations and evolving wind fields. Representative methods include CNN–Informer hybrids [9], graph convolutional network (GCN)–GRU models for offshore forecasting [10], and variational mode decomposition (VMD)-based CNN–GRU frameworks [11], all demonstrating improved accuracy through spatial–temporal coupling. Zhang et al. [12] introduced an ICEEMDAN–BiTCN–BiGRU model with attention for multi-scale feature extraction, while Ghanbari and Avar [13] developed an MVMD–LSTM framework for non-stationary signal prediction. Liu et al. [14] and Shringi et al. [15] further combined decomposition and attention-based recurrent models, achieving accurate multi-step forecasts.

The Transformer architecture has recently gained attention for long-range temporal modeling. Informer [16] reduces computational cost via sparse attention, while Autoformer [17] introduces an autocorrelation mechanism to capture periodicity. Further variants, such as CNN–Informer [18] and decomposition-enhanced Autoformer [19], enhance long-horizon forecasting. Nevertheless, most Transformer-based approaches still emphasize temporal learning while neglecting explicit spatial correlations and struggling with highly volatile, non-stationary wind dynamics. Their high computational cost and limited interpretability also hinder deployment in real-world wind farms.

To address these challenges, this paper proposes a Graph-Guided Spatio-Temporal Autoformer (GSTAformer), which integrates graph neural networks (GNNs) with the Autoformer architecture for unified spatio-temporal learning. The proposed framework is evaluated on two benchmark datasets—Baidu SDWPF and GEFCom2012—covering diverse spatio-temporal scales. Experimental results show that GSTAformer consistently outperforms state-of-the-art methods across multiple forecasting horizons (6, 12, and 24 h), achieving superior accuracy, robustness, and efficiency.

The main contributions are summarized as follows:

(1) Graph-guided Autoformer framework: A unified architecture integrating GNN-based spatial modeling with Autoformer, enabling joint learning of spatial and temporal dependencies for wind power forecasting.
(2) Multi-scale temporal fusion: A convolution-enhanced temporal module that captures both rapid fluctuations and periodic variations, improving adaptability to non-stationary wind dynamics.
(3) Comprehensive validation: Extensive experiments on SDWPF and GEFCom2012 confirm GSTAformer's superior performance across different forecasting horizons.

The remainder of this paper is organized as follows: Section 2 introduces the architecture and methodology of the proposed GSTAformer. Section 3 describes the datasets, experimental setup, and main results. Section 4 provides a detailed discussion of the findings, model behavior, and limitations. Finally, Section 5 concludes the paper and outlines future research directions.

2. Methodology

2.1. Overall Structure of the Proposed Network

To enhance medium-term wind power forecasting, we propose GSTAformer, a unified framework that jointly models spatial and temporal dependencies. As illustrated in Figure 1, the architecture comprises three components: (1) a spatial feature extraction module, (2) a temporal feature extraction module, and (3) an improved Autoformer.

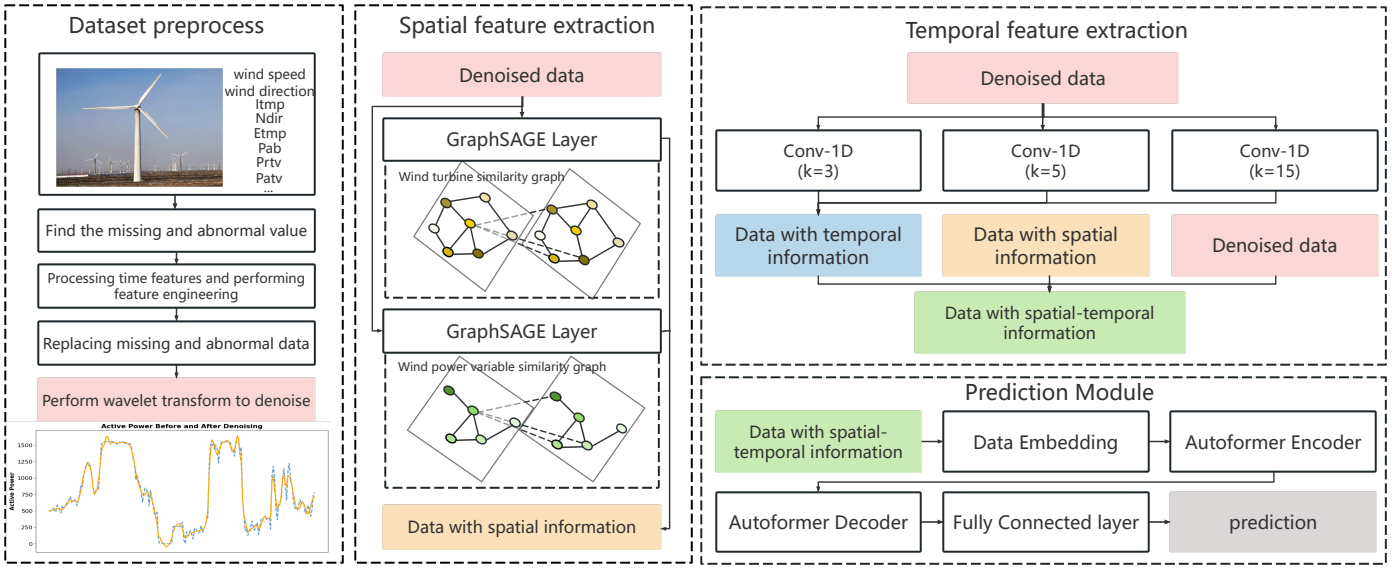


Figure 1. The structure of the GSTAformer for wind power forecasting. The blue line denotes the wind power series before wavelet-based denoising, whereas the yellow line denotes the denoised wind power series.

Preprocessed wind power data—after missing-value interpolation, anomaly filtering, and noise reduction via discrete wavelet transform (DWT) [20]—are first fed into the spatial module. Two graphs are constructed: a turbine-level graph using the maximal information coefficient (MIC) and a variable-level graph using the Pearson correlation coefficient (PCC). GraphSAGE then aggregates node features to obtain spatial embeddings.

These embeddings are processed by the temporal module, which employs multi-scale convolutions to capture both short-term variations and long-term trends. Finally, the fused spatio-temporal features are fed into the enhanced Autoformer with self-attention for accurate medium-term wind power forecasting.

2.2. Spatial Feature Extraction Module

Modeling spatial dependencies is crucial for wind power forecasting, as turbine outputs are strongly correlated by geography and meteorology. Many existing models use simple distance-based metrics or ignore inter-variable relations, leading to incomplete spatial representations. To address this, we design a dual-graph spatial feature extraction module leveraging both turbine-level and variable-level similarity graphs, followed by GraphSAGE aggregation to produce robust spatial embeddings as shown in Figure 2.

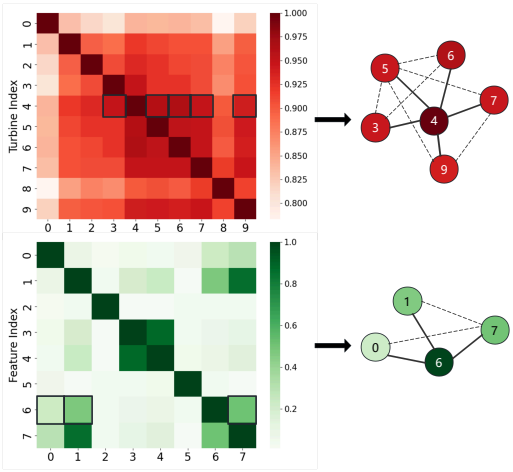


Figure 2. Dual-graph construction using MIC and PCC similarity matrices. The numerical values in the heatmaps represent the pairwise similarity coefficients computed by MIC and PCC, respectively.

2.2.1. Graph Construction

Turbine similarity graph. An undirected weighted graph $G_1 = (V_1, E_1)$ is built, where each node v_i denotes a turbine and edge weights encode nonlinear correlations of power series. Instead of Euclidean distance or PCC, we employ the maximal information coefficient (MIC) [21], which captures both linear and nonlinear dependencies:

$$W_{ij} = \mathrm{MIC}(x_i, x_j). \tag{1}$$

A high MIC value indicates strong shared information across operating conditions, which typically arises when turbines are aligned with dominant wind directions or influenced by similar terrain-induced flow structures. Thus, G_1 approximates flow connectivity rather than mere Euclidean proximity.

For each turbine v_i, only the k most similar neighbors are retained, forming the adjacency matrix $A \in \mathbb{R}^{m \times m}$:

$$A_{ij} = \begin{cases} W_{ij}, & v_j \in N(i) \text{ or } v_i \in N(j), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Variable correlation graph. To model inter-variable dependencies among meteorological and operational factors, an undirected weighted graph $G_2 = (V_2, E_2)$ is constructed using the Pearson correlation coefficient (PCC) [22]:

$$r_{ij} = \frac{\sum_{t=1}^{n}(x_i^t - \bar{x}_i)(x_j^t - \bar{x}_j)}{\sqrt{\sum_{t=1}^{n}(x_i^t - \bar{x}_i)^2}\sqrt{\sum_{t=1}^{n}(x_j^t - \bar{x}_j)^2}}. \tag{3}$$

Each variable node connects to its q most correlated neighbors, yielding adjacency matrix B. This complementary design captures nonlinear turbine similarities through G_1 and linear variable correlations through G_2, enabling comprehensive spatial dependency learning. The PCC-based graph G_2 captures physically consistent co-variations among variables such as wind speed, direction, and temperature, which are jointly modulated by mesoscale atmospheric regimes or operational strategies.

2.2.2. GraphSAGE-Based Feature Aggregation

Given G_1 and G_2, we employ GraphSAGE [23] to aggregate spatial features in an inductive manner, allowing generalization to unseen turbines or variables. Specifically, the hidden representation of node i is updated as

$$h_i' = \sigma(W_1 h_i + W_2 \cdot \mathrm{AGG}\{h_j \mid j \in N(i)\}), \tag{4}$$

where $\mathrm{AGG}(\cdot)$ denotes a neighborhood aggregation function (mean pooling in this study), and $\sigma(\cdot)$ is a nonlinear activation. The same process is applied to variable nodes f_i in G_2:

$$f_i' = \sigma(U_1 f_i + U_2 \cdot \mathrm{AGG}\{f_j \mid j \in M(i)\}). \tag{5}$$

To better capture inter-variable dependencies, we adopt the variable-as-token strategy inspired by iTransformer [24], and the contrast between the traditional GNN input view and the proposed view in GSTAformer is illustrated in Figure 3. Specifically, the input tensor $X \in \mathbb{R}^{m \times T \times C}$ is reshaped into $X_{\text{variable-major}} \in \mathbb{R}^{C \times m \times T}$, treating variables as tokens while preserving turbine and temporal dimensions. This design facilitates cross-variable interactions during aggregation.

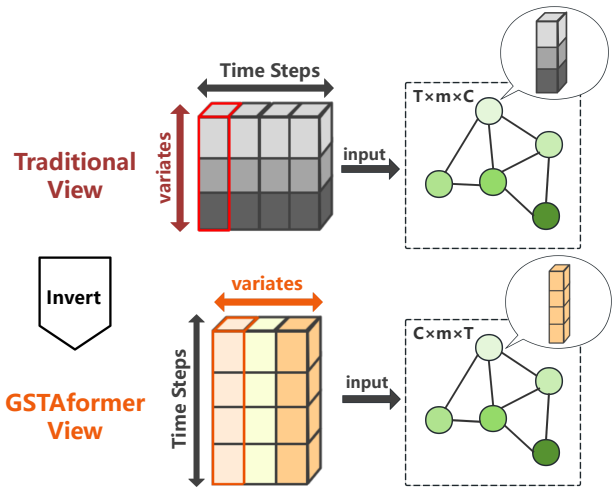


Figure 3. Traditional vs. GSTAformer view on graph neural network (GNN) input.

Finally, we obtain two sets of embeddings: turbine-level features $H_{\text{spatial}} \in \mathbb{R}^{m \times T \times C}$ from G_1, and variable-level features $H_{\text{feature}} \in \mathbb{R}^{m \times T \times C}$ from G_2, which are subsequently fused in the temporal module for spatio-temporal learning.

2.3. Temporal Feature Extraction Module

Wind power series exhibit multi-scale temporal variability, containing long-term periodicity (e.g., diurnal/seasonal cycles) and short-term fluctuations (e.g., gusts and turbulence). To capture both global and local dependencies beyond the capacity of recurrent models (LSTM, GRU), we design a multi-scale convolution (MSC) module for hierarchical temporal feature extraction.

As shown in Figure 4, the MSC module comprises three parallel one-dimensional convolutions with kernel sizes $k_1 = 3$, $k_2 = 5$, and $k_3 = 15$, corresponding to short-, medium-, and long-range dependencies. These kernel sizes are empirically determined from the autocorrelation structure of wind power data to balance responsiveness and smoothness. To verify robustness, we conduct a sensitivity experiment by replacing the kernel triplet with $\{3, 5, 7\}$ and $\{3, 7, 15\}$.

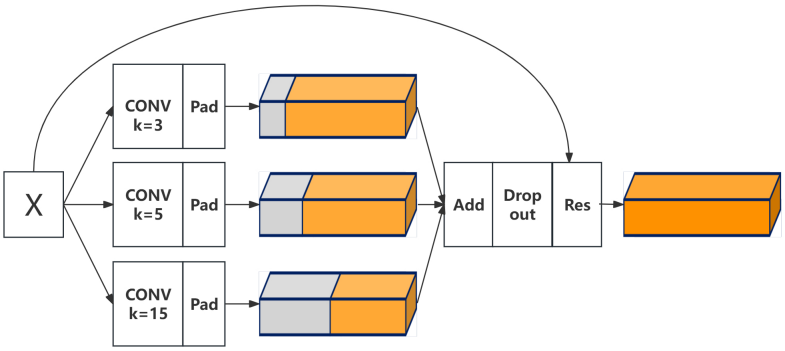


Figure 4. Flowchart of the temporal feature extraction module.

Given input $X \in \mathbb{R}^{m \times T \times C}$, the ith branch performs

$$F_i = \phi(W_i * X + b_i), \tag{6}$$

where $*$ denotes temporal convolution, and $\phi(\cdot)$ is the rectified linear unit (ReLU). Zero-padding preserves dimensional consistency. The three feature maps are fused by element-wise summation:

$$F_{\text{time}} = \hat{F}_1 + \hat{F}_2 + \hat{F}_3, \tag{7}$$

yielding a unified representation that integrates short-term dynamics and long-term trends for subsequent modeling.

2.4. Improved Autoformer

To enhance long-term forecasting and strengthen the integration of spatial–temporal representations, we redesign the Autoformer module with two refinements: (1) an adaptive decomposition window and (2) a full-attention mechanism in the decoder as illustrated in Figure 5.

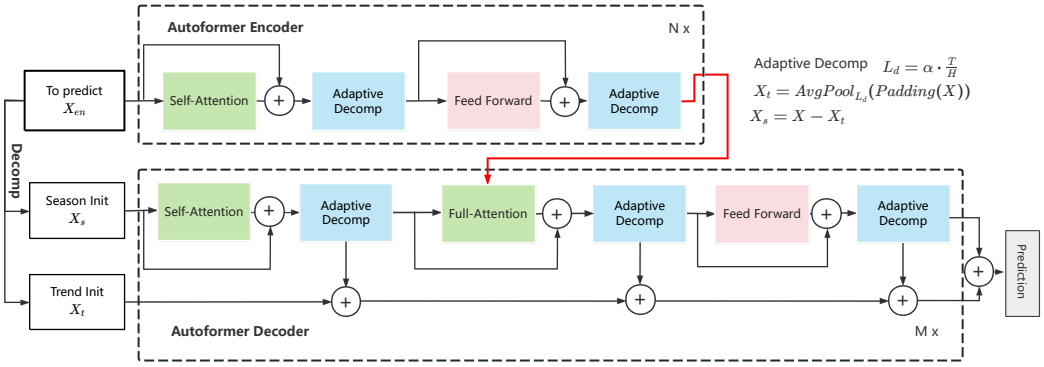


Figure 5. Structure of the improved Autoformer.

2.4.1. Adaptive Series Decomposition

The original Autoformer [17] employs a fixed moving-average window for trend–seasonal decomposition, which limits its ability to adapt to non-stationary wind sequences. We introduce an adaptive decomposition window whose length L_d varies with the input length T and prediction horizon H:

$$L_d = \alpha \cdot \frac{T}{H}, \tag{8}$$

where α is a tunable factor ($\alpha = 0.5$ by default). To justify this choice, we perform a sensitivity analysis using $\alpha \in \{0.25, 0.50, 0.75\}$. A shorter window preserves rapid variations,

while a longer one extracts smooth long-term dynamics. This operator decomposes the sequence into trend and seasonal parts:

$$X_t = \text{AvgPool}_{L_d}\big(\text{Padding}(X)\big), \tag{9}$$

$$X_s = X - X_t, \tag{10}$$

where X_t captures slowly evolving patterns, and X_s models short-term fluctuations. The adaptive window improves robustness across diverse temporal scales.

2.4.2. Full-Attention Mechanism

To enhance global dependency modeling, we replace conventional cross-attention in the decoder with a full-attention mechanism, allowing each decoder query to attend jointly to encoder outputs and its own autoregressive history. Let H_{enc} denote the encoder output and H_{dec} the decoder input. We form a unified memory bank

$$M = \text{Concat}(H_{\text{enc}}, H_{\text{dec}}), \tag{11}$$

from which keys and values are generated. Full-attention is computed as

$$\text{FullAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{12}$$

where causal masking is applied to preserve autoregressive ordering. This design enables the decoder to fuse long-term historical context with evolving seasonal patterns, improving representation completeness over extended forecasting horizons.

2.4.3. Encoder–Decoder Processing

Each encoder layer refines seasonal representations through self-attention, feed-forward processing, and adaptive decomposition:

$$S_{\text{en}}^{(l,1)} = \text{SeriesDecomp}\Big(\text{SelfAttn}(X_{\text{en}}^{(l-1)}) + X_{\text{en}}^{(l-1)}\Big), \tag{13}$$

$$S_{\text{en}}^{(l,2)} = \text{SeriesDecomp}\Big(\text{FFN}(S_{\text{en}}^{(l,1)}) + S_{\text{en}}^{(l,1)}\Big), \tag{14}$$

ensuring that long-range temporal dependencies are extracted while gradually removing accumulated trend bias.

In the decoder, historical seasonal components are extended with zero-initialized future positions, and trend components are initialized using historical statistics to stabilize extrapolation. Each decoding layer combines self-attention and full-attention to incorporate both short-term seasonal structure and global encoder memory:

$$S_{\text{de}}^{(l,2)},\ T_{\text{de}}^{(l,2)} = \text{SeriesDecomp}\Big(\text{FullAttn}(S_{\text{de}}^{(l,1)},\ H_{\text{enc}}^{N}) + S_{\text{de}}^{(l,1)}\Big), \tag{15}$$

$$T_{\text{de}}^{(l)} = T_{\text{de}}^{(l-1)} + T_{\text{de}}^{(l,2)}. \tag{16}$$

Seasonal and trend predictions are finally aggregated to obtain the multi-step forecast. The improved Autoformer therefore adaptively decomposes temporal signals and strengthens long-range dependency learning for complex wind dynamics.

3. Experiment

3.1. Data Description

3.1.1. SDWPF Dataset

The SDWPF dataset from the KDD Cup 2022 contains 10 min records from 134 turbines over 245 days, including wind speed, direction, temperature, and power output. Ten representative turbines are selected (about 35,000 samples each) to provide diverse temporal and spatial characteristics. Data quality control follows turbine-level physical constraints: measurements with negative power, zero power under sufficiently high wind speed, abnormal blade pitch angles, or out-of-range wind direction are treated as invalid and replaced with missing values. Missing segments are imputed using backward–forward filling, which is suitable for the smooth evolution of turbine operating states. Feature engineering includes normalized temporal variables (day, hour, and minute), absolute rotor speed, and the maximum blade pitch angle as a stability indicator, while irrelevant columns are removed.

3.1.2. GEFCom2012 Dataset

The GEFCom2012 dataset provides hourly normalized wind power data from seven wind farms (2009–2012) with 48 h forecasting tasks. It also includes data gaps caused by maintenance and sensor faults, yielding realistic testing conditions. Since no turbine-level measurements are available, only calendar-based features (Year/Month/Day/Hour) are extracted and normalized to $[-0.5, 0.5]$, and dataset-specific auxiliary fields are removed to avoid redundancy.

Both datasets are divided into training, validation, and test sets with a 7:1:2 ratio. Each experiment is repeated ten times and averaged for stability. Inputs are min–max normalized, and the input length is set to twice the prediction horizon to ensure sufficient historical context. To suppress sensor noise while preserving ramp dynamics important for forecasting, wavelet denoising is applied to power sequences using a Daubechies-4 wavelet, two decomposition levels, and soft-thresholding based on the MAD-estimated universal threshold.

3.2. Comparison with Baseline Models

To validate the effectiveness of GSTAformer, we compare it with 13 representative baselines on the SDWPF and GEFCom2012 datasets, covering three model categories:

RNN-based models: LSTM, GRU, bidirectional LSTM (BiLSTM), and bidirectional GRU (BiGRU), representing classical sequence models for temporal dependency learning.

CNN-based model: Temporal convolutional network (TCN) [25], which captures temporal patterns via dilated causal convolutions.

Transformer-based models: DLinear [26], Transformer [27], Informer [16], PatchTST [28], iTransformer [24], and TimeXer [29], representing recent advances in long-sequence forecasting.

All baselines are trained under identical data splits, input/output lengths, normalization, and optimization settings with early stopping for fair comparison.

3.2.1. Evaluation on the SDWPF Dataset

Table 1 summarizes the forecasting results on the SDWPF dataset. GSTAformer consistently achieves the lowest MSE and root-mean-square error (RMSE) across all horizons, confirming its superior capacity to capture both short-term fluctuations and long-term temporal trends. Figure 6 further illustrates its robustness.

Table 1. Comprehensive evaluation of model performance on the SDWPF dataset.

Model	6-h			12-h			24-h		
	MSE	**MAE**	**RMSE**	**MSE**	**MAE**	**RMSE**	**MSE**	**MAE**	**RMSE**
LSTM	0.4253	0.4285	0.6521	0.6907	0.5858	0.8308	0.8789	0.6999	0.9375
GRU	0.4070	0.4213	0.6379	0.6538	0.5832	0.8086	0.8866	0.7100	0.9415
BiLSTM	0.4441	0.4657	0.6664	0.6716	0.5845	0.8194	0.9164	0.7078	0.9571
BiGRU	0.4273	0.4387	0.6537	0.6717	0.6002	0.8194	0.8921	0.7229	0.9444
TCN [25]	0.4096	0.4408	0.6400	0.6344	0.5773	0.7964	0.8657	0.6896	0.9304
DLinear [26]	0.4210	0.4779	0.6488	0.6434	0.6121	0.8021	0.8462	0.7298	0.9199
Transformer [27]	0.4933	0.4609	0.7009	0.8826	0.6503	0.9390	1.2029	0.7949	1.0938
Informer [16]	0.5179	0.4759	0.7192	0.8135	0.6254	0.9018	1.1274	0.7595	1.0583
PatchTST [28]	0.4123	**0.3983**	0.6420	0.7361	0.5888	0.8579	1.0648	0.7249	1.0303
iTransformer [24]	0.4574	0.4204	0.6762	0.7796	0.5835	0.8828	1.1849	0.7480	1.0867
TimeXer [29]	0.4555	0.4259	0.6747	0.7488	0.5844	0.8653	1.0731	0.7216	1.0344
GSTAformer	**0.3887**	0.4210	**0.6234**	**0.5469**	**0.5364**	**0.7395**	**0.7480**	**0.6362**	**0.8648**

Note: Bold values indicate the best performance for each prediction horizon.

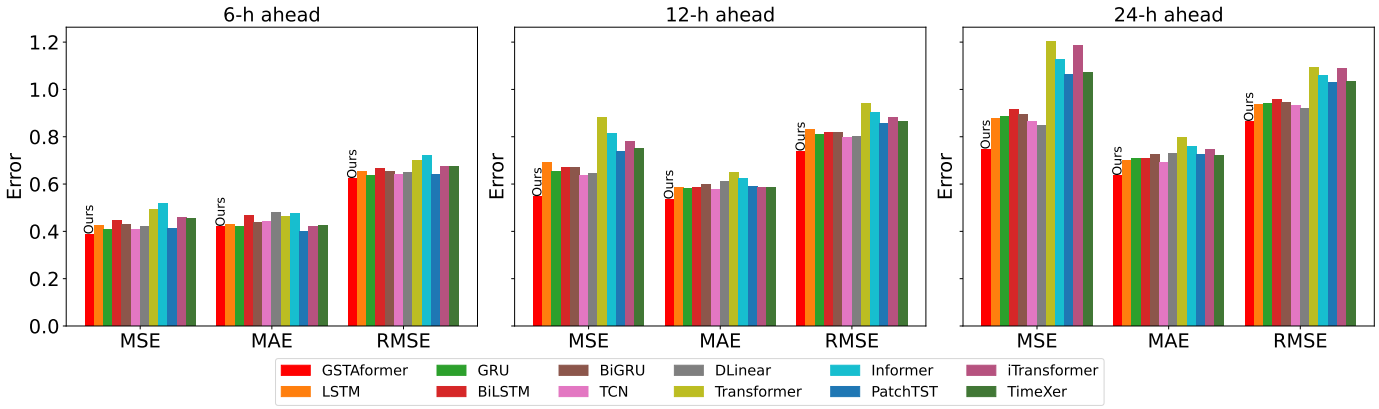


Figure 6. Performance comparison of forecasting models on the SDWPF dataset.

6 h Ahead Forecasting

As shown in Table 1, short-term forecasting yields relatively small errors across models. GSTAformer attains the best MSE (0.3887) and RMSE (0.6234), outperforming all baselines. The improvement arises from the MSC module, which extracts fine-grained local variations that recurrent models tend to over-smooth. Although PatchTST achieves a slightly lower MAE, its higher MSE and RMSE indicate less stability under fluctuating conditions.

12 h Ahead Forecasting

For the 12 h horizon, increased temporal dependency makes forecasting more challenging. GSTAformer achieves the lowest MSE (0.5469), MAE (0.5364), and RMSE (0.7395), improving upon TCN by 13.81% in MSE. These gains stem from the GNN-based spatial integration and the adaptive series decomposition, which separates seasonal and trend components for more accurate mid-range prediction.

24 h Ahead Forecasting

Long-horizon prediction is most affected by uncertainty and trend drift. GSTAformer still leads with MSE = 0.7480, MAE = 0.6362, and RMSE = 0.8648. The full-attention mechanism and adaptive decomposition window jointly enhance the capture of global periodicity while preserving high-frequency dynamics. Compared with LSTM and GRU, MSE decreases by 14.9% and 15.7%, respectively, verifying the effectiveness of the spatio-temporal fusion strategy.

3.2.2. Evaluation on the GEFCom2012 Dataset

Table 2 reports results on the GEFCom2012 dataset, and Figure 7 provides a visual comparison of the forecasting performance across different models and horizons. All models perform better than on SDWPF due to lower volatility and stronger periodicity, yet GSTAformer remains best across all horizons, highlighting its strong generalization ability.

Table 2. Comprehensive evaluation of model performance on the GEFCom2012 dataset.

Model	6-h			12-h			24-h		
	MSE	**MAE**	**RMSE**	**MSE**	**MAE**	**RMSE**	**MSE**	**MAE**	**RMSE**
LSTM	0.3294	0.3413	0.5740	0.4177	0.3913	0.6463	0.4910	0.4292	0.7007
GRU	0.3401	0.3509	0.5832	0.4201	0.3910	0.6481	0.4943	0.4310	0.7031
BiLSTM	0.3128	0.3289	0.5593	0.4054	0.3844	0.6367	0.4806	0.4236	0.6933
BiGRU	0.3263	0.3388	0.5712	0.4235	0.3927	0.6507	0.4845	0.4251	0.6961
TCN [25]	0.3359	0.3523	0.5795	0.4148	0.3940	0.6440	0.5091	0.4509	0.7135
DLinear [26]	0.3106	0.3205	0.5572	0.4037	0.3712	0.6353	0.5227	0.4552	0.7230
Transformer [27]	0.2780	0.2760	0.5266	0.3984	0.3377	0.6303	0.5389	0.4354	0.7336
Informer [16]	0.2988	0.3017	0.5462	0.4113	0.3670	0.6399	0.5332	0.4571	0.7298
PatchTST [28]	0.2332	0.2310	0.4828	0.3478	0.3239	0.5897	0.4406	0.4017	0.6636
iTransformer [24]	0.2327	0.2306	0.4823	0.3330	0.3282	0.5769	0.4747	0.3925	0.6890
TimeXer [29]	0.2176	0.2303	0.4665	0.3431	0.3260	0.5856	0.4522	0.4104	0.6724
GSTAformer	**0.1958**	**0.2297**	**0.4425**	**0.3080**	**0.3207**	**0.5549**	**0.4001**	**0.3799**	**0.6325**

Note: Bold values indicate the best performance for each prediction horizon.

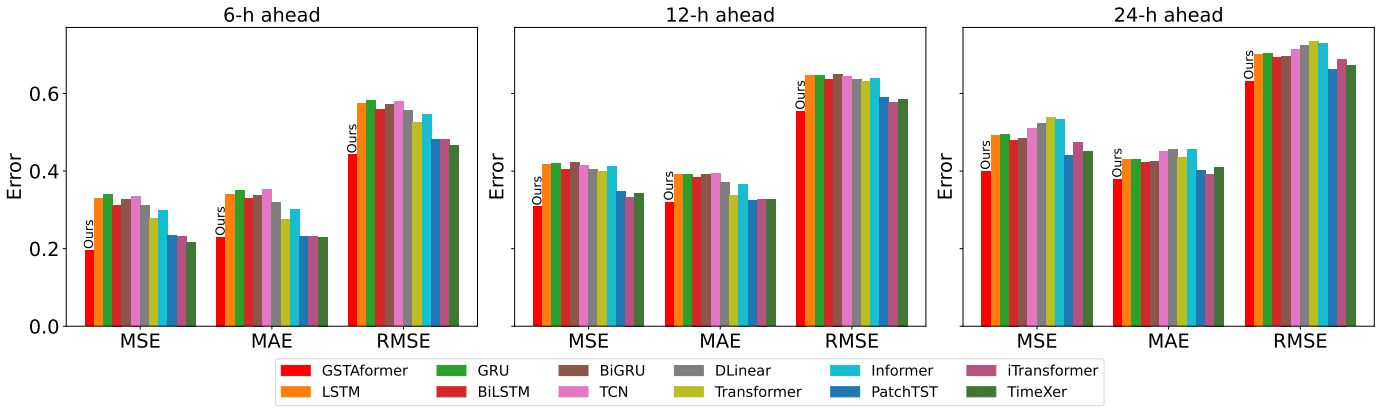


Figure 7. Performance comparison of forecasting models on the GEFCom2012 dataset.

6 h Ahead Forecasting

GSTAformer achieves the lowest MSE (0.1958), MAE (0.2297), and RMSE (0.4425), outperforming TimeXer by 10.03% in MSE. The multi-scale convolution module captures fine-grained patterns that characterize this smoother dataset.

12 h Ahead Forecasting

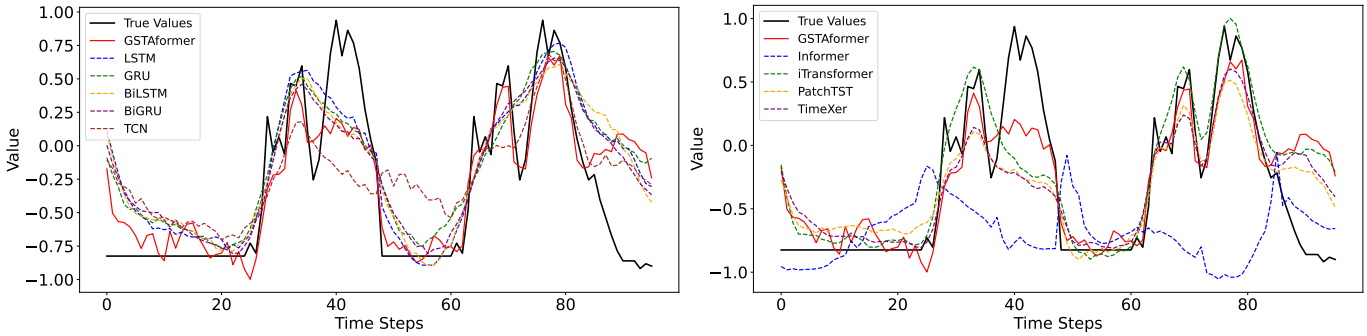
For the 12 h horizon, GSTAformer further improves performance through its adaptive decomposition window, which adjusts to the dataset's periodicity. Compared with iTransformer, MSE and RMSE are reduced by 7.51% and 3.81%, respectively.

24 h Ahead Forecasting

At the 24 h horizon, GSTAformer attains MSE = 0.4001, MAE = 0.3799, and RMSE = 0.6325, demonstrating consistent trend tracking and short-term adaptability. PatchTST and TimeXer show weaker responses to anomalies, whereas GSTAformer effectively balances spatial and temporal dependencies for more accurate long-term forecasts.

Further Analysis

Figure 8 shows predictive trajectories. GSTAformer more accurately tracks sharp transitions (steps 30–50 and 60–80), reflecting its adaptability to dynamic variations. The synergy between GNN-based spatial representations and full-attention temporal modeling mitigates long-horizon error accumulation.



(**a**) Traditional model (**b**) Latest Model

Figure 8. Diagram of forecasting result curves on the GEFCom2012 dataset.

3.3. Ablation Study

To further evaluate the contribution of each component, we conduct comprehensive ablation experiments on the SDWPF dataset, as shown in Table 3. All variants are trained under identical hyperparameters and parameter scales to ensure fair comparison, and results are averaged over five runs for statistical stability (standard deviation within ±1.3%).

The evaluated variants include the following:

- w/o GTM: removing the graph–temporal fusion module, which jointly integrates spatial and temporal embeddings;
- w/o Temporal: excluding the MSC-based temporal feature extractor;
- w/o G_1: removing the turbine-level spatial graph modeling component;
- w/o G_2: removing the variable-level correlation graph module.

Table 3. Ablation study results of GSTAformer on multiple forecasting scales (SDWPF dataset).

Model	6-h			12-h			24-h		
	MSE	**MAE**	**RMSE**	**MSE**	**MAE**	**RMSE**	**MSE**	**MAE**	**RMSE**
GSTAformer w/o GTM	0.4120	0.4247	0.6418	0.6391	0.5692	0.7992	0.8609	0.6892	0.9274
GSTAformer w/o Temporal	0.3990	0.4318	0.6315	0.5483	0.5397	0.7395	0.7651	0.6508	0.8744
GSTAformer w/o G_1	0.3978	0.4275	0.6307	0.5967	0.5512	0.7716	0.7707	0.6570	0.8775
GSTAformer w/o G_2	0.3908	0.4281	0.6251	0.5677	0.5440	0.7525	0.8265	0.6827	0.9084
GSTAformer	**0.3887**	**0.4210**	**0.6234**	**0.5469**	**0.5364**	**0.7395**	**0.7480**	**0.6362**	**0.8648**

Note: Bold values indicate the best performance for each prediction horizon.

3.3.1. Spatial–Temporal Coupling Effect

Removing the graph–temporal fusion module (w/o GTM) causes the largest performance degradation across all horizons, particularly in the 24 h forecasting task where MSE increases from 0.7480 to 0.8609. This confirms that the integrated spatio-temporal representation is indispensable for long-horizon forecasting, as it allows dynamic interaction between turbine-level topology and evolving temporal dynamics. Eliminating the turbine-level graph (w/o G_1) also notably worsens results, with MSE rising from 0.5469 to 0.5967 in the 12 h setting, highlighting the importance of explicitly modeling turbine interdependence. In contrast, removing the variable-level graph (w/o G_2) produces a smaller

but consistent decline, suggesting that feature-level correlation contributes to smoother local temporal prediction.

The different behaviors of G_1 and G_2 across horizons can be explained as follows. For 6 h and 12 h forecasts, removing G_1 leads to a larger MSE increase than removing G_2 (e.g., from 0.5469 to 0.5967 at 12 h) because these horizons are dominated by the short- to mid-term propagation of wind fields along the turbine layout, which is explicitly captured by the MIC-based turbine graph. At 24 h, however, the absence of G_2 is more detrimental (MSE 0.8265 vs. 0.7707 for w/o G_1), indicating that very long-range prediction relies more on stable cross-variable relationships (e.g., between wind speed, direction and power) encoded in the PCC-based graph to regularize the temporal evolution. This suggests that G_1 mainly enhances short- to medium-term responsiveness, whereas G_2 provides long-term stabilization, and their combination yields the best overall accuracy.

3.3.2. Temporal Dynamics Effect

Excluding the multi-scale temporal feature extractor (w/o Temporal) increases MSE from 0.7480 to 0.7651 in the 24 h horizon, indicating that the MSC module effectively captures fine-grained short- and medium-term dynamics. Without this component, the model struggles to represent local fluctuations and fails to align temporal patterns across multiple scales.

3.4. Hyperparameter Sensitivity Analysis

To evaluate the robustness of key architectural design choices in GSTAformer, we conduct two sensitivity experiments on the SDWPF dataset for the 24 h horizon. These experiments examine the effect of (1) the MSC kernel-size configuration and (2) the decomposition coefficient α used in the adaptive trend extraction.

Table 4 compares three kernel-size triplets. Although all configurations produce comparable errors, the proposed setting $\{3, 5, 15\}$ achieves the best performance (MSE = 0.7480), indicating its balanced ability to capture short-, medium-, and long-range temporal patterns.

Table 5 reports results for $\alpha \in \{0.25, 0.50, 0.75\}$. The model performs best at $\alpha = 0.5$, while nearby values yield slightly higher errors (0.7530–0.7600 MSE), demonstrating that GSTAformer is not overly sensitive to this hyperparameter.

These results confirm that the proposed architectural settings are well motivated and stable across reasonable hyperparameter ranges.

Table 4. Sensitivity of MSC kernel sizes on the SDWPF dataset (24 h horizon).

Model (Kernel Sizes)	MSE (24-h)	MAE (24-h)	RMSE (24-h)
GSTAformer ($k = \{3, 5, 7\}$)	0.7580	0.6420	0.8706
GSTAformer ($k = \{3, 5, 15\}$)	0.7480	0.6362	0.8648
GSTAformer ($k = \{3, 7, 15\}$)	0.7540	0.6395	0.8683

Table 5. Sensitivity of the adaptive decomposition coefficient α on the SDWPF dataset (24 h horizon).

Model (α)	MSE (24-h)	MAE (24-h)	RMSE (24-h)
GSTAformer ($\alpha = 0.25$)	0.7600	0.6430	0.8718
GSTAformer ($\alpha = 0.50$)	0.7480	0.6362	0.8648
GSTAformer ($\alpha = 0.75$)	0.7530	0.6390	0.8678

3.5. Computational Cost Analysis

To assess the efficiency and feasibility of GSTAformer in practical wind farm operations, we evaluate the computational cost on the SDWPF dataset under the 72-step–to–36-step forecasting setting, which is one of the most computationally demanding configu-

rations. We report the number of parameters, FLOPs per forward pass, and per-epoch training time, measured on a single NVIDIA GeForce RTX 4090 GPU.

Table 6 shows the comparison results. GSTAformer achieves a favorable balance between efficiency and modeling capability. With 3.17M parameters and 42.48 GFLOPs per forward pass, its computational footprint is significantly lower than that of mainstream Transformer models, while its training speed (55.56 s per epoch) remains well within the range suitable for large-scale, continuous retraining in operational wind farms.

Table 6. Computational cost on the SDWPF dataset (72→36 forecasting).

Model	Params (M)	FLOPs (GFLOPs)	Time/Epoch (s)
LSTM	1.088	12.366	16.17
GRU	0.820	9.281	11.97
BiLSTM	2.175	24.731	32.34
BiGRU	1.640	18.562	24.27
TCN	0.255	3.085	4.03
DLinear	0.006	0.015	0.02
Transformer	10.539	121.333	160.86
Informer	11.327	110.082	143.96
PatchTST	6.480	72.902	95.34
iTransformer	6.362	11.188	14.63
TimeXer	8.547	6.966	9.11
GSTAformer	3.166	42.484	55.56

4. Discussion

The experiments on the SDWPF and GEFCom2012 datasets show that GSTAformer consistently outperforms recurrent, convolutional, and Transformer-based baselines over 6-, 12-, and 24 h horizons (Tables 1 and 2). On the more volatile SDWPF dataset, the gains are most pronounced for 24 h forecasting, suggesting that explicit spatio-temporal modeling effectively mitigates long-horizon error accumulation. On the smoother and more periodic GEFCom2012 dataset, GSTAformer still achieves the best performance, indicating good generalization across different temporal characteristics and spatial layouts.

From a practical perspective, the two case studies considered in this work already represent distinct operational regimes: SDWPF corresponds to a highly volatile wind farm with fine temporal resolution, whereas GEFCom2012 represents aggregated wind power with smoother, more periodic behavior. The consistent gains of GSTAformer across these datasets suggest that the proposed architecture can generalize to different wind farm configurations, aggregation levels, and temporal resolutions, which is essential for deployment in heterogeneous power systems.

4.1. Effect of Spatial Graph Modeling

The ablation study highlights the importance of the dual-graph design. Removing the turbine-level similarity graph (w/o G_1) leads to clear degradation, especially at 12 h and 24 h, confirming that MIC-based turbine correlations provide useful inductive bias beyond purely temporal modeling (Table 3). The variable-level correlation graph (G_2) also brings consistent, though slightly smaller, improvements, suggesting that encoding relationships among meteorological and operational variables helps regularize feature learning. This interpretation is consistent with the quantitative differences reported in Table 3, where removing either G_1 or G_2 systematically increases MSE, MAE, and RMSE across all horizons, and the full model with both graphs always yields the best performance.

4.2. Role of Multi-Scale Temporal Modeling and Improved Autoformer

The multi-scale convolution (MSC) module and the improved Autoformer jointly strengthen temporal modeling. Ablation results show that removing the MSC module (w/o Temporal) particularly harms longer-horizon performance on SDWPF, indicating that multi-scale convolutions complement attention by capturing fine-grained local dynamics that are easily smoothed out in purely global architectures. The adaptive decomposition window allows the trend–seasonal separation to adjust to different input and prediction lengths, improving robustness to non-stationary behavior. Meanwhile, the full-attention mechanism, where decoder queries attend to both encoder outputs and their own history, enhances long-range dependency modeling. Compared with the vanilla Autoformer, GSTAformer achieves superior accuracy with only modest increases in parameters and training time, suggesting a favorable trade-off between effectiveness and efficiency.

Furthermore, the kernel-size and decomposition-coefficient sensitivity experiments (Tables 4 and 5) show that the model maintains its advantage under reasonable variations of these temporal-design hyperparameters. This indicates that the improvements brought by the MSC and adaptive decomposition modules are robust rather than the result of fragile hyperparameter tuning, thereby providing stronger support for the architectural choices adopted in GSTAformer.

4.3. Limitations and Future Directions

Despite its strong performance, GSTAformer has several limitations. The dual-graph and multi-branch design increases architectural complexity and may limit interpretability and deployment on resource-constrained platforms. Moreover, the current spatial graphs are constructed from historical correlations and remain static during inference, which ensures stability and computational efficiency but cannot fully capture the evolving spatial dependencies induced by changing atmospheric conditions. The present work also focuses only on deterministic point forecasting and does not provide uncertainty estimation, which is essential for risk-aware decision-making in power dispatch and electricity trading.

Future work will explore dynamic or learned graph structures that adapt to evolving wind field patterns in an online manner, as well as model compression and acceleration techniques to further reduce computational cost. In addition, extending GSTAformer to probabilistic forecasting through distributional outputs, Monte Carlo sampling, or ensemble schemes would enable quantification of predictive uncertainty.

5. Conclusions

This paper presented GSTAformer, a graph-guided spatio-temporal forecasting framework that integrates (i) a dual-graph architecture for learning local and global spatial dependencies, (ii) a multi-scale temporal fusion module for capturing both rapid fluctuations and periodic trends, and (iii) a unified Autoformer-based pipeline that jointly models spatial–temporal interactions.

Extensive experiments on two representative operational regimes—SDWPF and GEFCom2012—demonstrate consistent and substantial improvements over 13 strong baselines. Across 6–24 h forecasting horizons, GSTAformer reduces prediction errors by roughly 10–20% compared with the best existing models, confirming its robustness and strong generalization capability under both highly volatile and more periodic wind conditions.

While promising, the framework introduces additional architectural complexity and relies on static spatial graphs. Future work will explore adaptive or learned graph structures, model compression for deployment efficiency, and probabilistic extensions to enable uncertainty-aware forecasting in real-world power system operations.

Author Contributions: Conceptualization, S.Y., Y.M., C.T. and M.X.; methodology, S.Y., Y.M., T.G. and F.Y.; software, S.Y.; validation, T.G., C.T. and F.Y.; formal analysis, T.G., C.T. and F.Y.; investigation, S.Y., Y.M. and M.X.; resources, M.X.; data curation, S.Y.; writing—original draft preparation, S.Y.; writing—review and editing, M.X.; visualization, Y.M.; supervision, M.X.; project administration, M.X.; and funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Project of SGCC, named Research on AI Trustworthy Assessment Technology for Power System Dispatching in Typical Scenarios (5108-202417039A-1-1-ZN).

Data Availability Statement: The original data presented in the study are openly available in FigShare at the following DOIs: SDWPF dataset (https://doi.org/10.6084/m9.figshare.30787586) and GEFCom2012 dataset (https://doi.org/10.6084/m9.figshare.30787913).

Acknowledgments: The authors would like to thank the organizers of the KDD Cup 2022 and the Global Energy Forecasting Competition 2012 for providing access to the SDWPF and GEFCom2012 datasets.

Conflicts of Interest: Author Fang Yu was employed by the company China Electric Power Research Institute Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Bokde, N.; Feijóo, A.; Villanueva, D.; Kulat, K. A review on hybrid empirical mode decomposition models for wind speed and wind power prediction. *Energies* **2019**, *12*, 254. [CrossRef]
2. Li, C.; Xiao, Z.; Xia, X.; Zou, W.; Zhang, C. A hybrid model based on synchronous optimisation for multi-step short-term wind speed forecasting. *Appl. Energy* **2018**, *215*, 131–144. [CrossRef]
3. Zhang, Y.; Wang, J.; Wang, X. Review on probabilistic forecasting of wind power generation. *Renew. Sustain. Energy Rev.* **2014**, *32*, 255–270. [CrossRef]
4. Lei, X.; Zhong, J.; Chen, Y.; Shao, Z.; Jian, L. Grid Integration of Electric Vehicles within Electricity and Carbon Markets: A Comprehensive Overview. *eTransportation* **2025**, *25*, 100435. [CrossRef]
5. Hanifi, S.; Liu, X.; Lin, Z.; Lotfian, S. A critical review of wind power forecasting methods—Past, present and future. *Energies* **2020**, *13*, 3764. [CrossRef]
6. Feng, R.; Jiang, S.; Liang, X.; Xia, M. STGAT: Spatial–Temporal Graph Attention Neural Network for Stock Prediction. *Appl. Sci.* **2025**, *15*, 4315. [CrossRef]
7. Xiong, B.; Lou, L.; Meng, X.; Wang, X.; Ma, H.; Wang, Z. Short-term wind power forecasting based on Attention Mechanism and Deep Learning. *Electr. Power Syst. Res.* **2022**, *207*, 107776. [CrossRef]
8. Boucetta, L.N.; Amrane, Y.; Arezki, S. Wind power forecasting using a GRU attention model for efficient energy management systems. *Electr. Eng.* **2024**, *107*, 2595–2620. [CrossRef]
9. Wang, H.; Song, K.; Cheng, Y. A hybrid forecasting model based on CNN and informer for short-term wind power. *Front. Energy Res.* **2022**, *9*, 788320. [CrossRef]
10. Xu, Z.; Liu, Y.; Zhao, Y.; Zhang, Q. Short-term prediction of offshore wind speed using a hybrid GCN–GRU model based on spatial-temporal correlations. *Energy Rep.* **2023**, *9*, 7096–7110. [CrossRef]
11. Zhao, J.; Huang, L.; Zhou, W. Short-term wind power prediction using VMD and CNN-GRU hybrid model. *Appl. Energy* **2024**, *350*, 121553. [CrossRef]
12. Zhang, X.; Ye, J.; Gao, L.; Ma, S.; Xie, Q.; Huang, H. Short-term wind power prediction based on ICEEMDAN decomposition and BiTCN–BiGRU–multi-head self-attention model. *Electr. Eng.* **2025**, *107*, 2645–2662. [CrossRef]
13. Ghanbari, E.; Avar, A. Short-term wind power forecasting using the hybrid model of multivariate variational mode decomposition (MVMD) and long short-term memory (LSTM) neural networks. *Electr. Eng.* **2025**, *107*, 2903–2933. [CrossRef]
14. Liu, J.; Wu, Y.; Cheng, X.; Li, B.; Yang, P. Short-term wind power prediction based on ICEEMDAN–Correlation reconstruction and BWO–BiLSTM. *Electr. Eng.* **2025**, *107*, 1381–1396. [CrossRef]
15. Shringi, S.; Saini, L.M.; Aggarwal, S.K. Multi-step ahead wind power forecasting based on multi-feature wavelet decomposition and convolution-gated recurrent unit model. *Electr. Eng.* **2025**, *107*, 9445–9466. [CrossRef]
16. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *arXiv* **2020**, arXiv:2012.07436. [CrossRef]

17. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430. [CrossRef]
18. Gong, M.; Yan, C.; Xu, W.; Zhao, Z.; Li, W.; Liu, Y.; Li, S. Short-term wind power forecasting model based on temporal convolutional network and Informer. *Energy* **2023**, *283*, 129171. [CrossRef]
19. Ban, G.; Chen, Y.; Xiong, Z.; Zhuo, Y.; Huang, K. The univariate model for long-term wind speed forecasting based on wavelet soft threshold denoising and improved Autoformer. *Energy* **2024**, *290*, 130225. [CrossRef]
20. Mannelli, A.; Papi, F.; Pechlivanoglou, G.; Ferrara, G.; Bianchini, A. Discrete Wavelet Transform for the Real-Time Smoothing of Wind-Turbine Power Output. *Energies* **2021**, *14*, 2184. [CrossRef]
21. Reshef, D.N.; Reshef, Y.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef]
22. Pearson, K. Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* **1895**, *58*, 240–242. [CrossRef]
23. Hamilton, W.L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 1025–1035.
24. Li, Y.; Wang, X.; Liu, J.; Zhang, T.; Wu, Q.; Xu, K.; Yan, J. iTransformer: Revisiting Channel Independent Strategy for Time Series Forecasting. *arXiv* **2023**, arXiv:2310.10774.
25. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]
26. Zeng, A.; Zhang, X.; Zheng, L.; Yi, X.; Khandekar, R.; Wang, Y.; Zhang, Y. Are Transformers Effective for Time Series Forecasting? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 28800–28812. [CrossRef]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
28. Zhao, Z.; Liu, J.; Zhang, Y.; Zhang, X.; Li, J.; Liu, Y. An Empirical Evaluation of Transformer-based Models for Long-term Time Series Forecasting. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
29. Wang, Y.; Zhang, H.; Liu, S.; Xu, K. TimeXer: Enhancing Time Series Forecasting with Exogenous Information. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 9–15 December 2024.