Marine Energy Research

SCIEPublish

Article

MSTGraph-RAR: Multi-Step Offshore Wind Power Forecasting with Variable Graph Learning and Ramp-Aware Residual Correction

Mingju Gong [1], Linlin Zhu [1], Yu Tian [1], Siyuan Yin [1], Yaming Wang [1], Yilin Zhang [1], Wei Xu [2,3,*] and Fei Su [1,*]

[1] School of Integrated Circuit Science and Engineering, Tianjin University of Technology, Tianjin 300384, China; gmj790@email.tjut.edu.cn (M.G.); zhulinlin@stud.tjut.edu.cn (L.Z.); tianyu@stud.tjut.edu.cn (Y.T.); yinsiyuan@stud.tjut.edu.cn (S.Y.); wangyaming@stud.tjut.edu.cn (Y.W.); zyl2487191829@stud.tjut.edu.cn (Y.Z.)

[2] Sea Island Environment Science and Technology Research Institute Co., Ltd., Tianjin 300000, China

[3] School of Ocean Energy, Tianjin University of Technology, Tianjin 300384, China

* Corresponding author. E-mail: 13920037327@email.tjut.edu.cn (W.X.); sufei@email.tjut.edu.cn (F.S.)

Received: 2 September 2026; Revised: 10 October 2026; Accepted: 23 October 2026; Available online: 5 November 2026

ABSTRACT: Accurate offshore wind power forecasting is essential for grid integration and secure operation, yet pronounced non-stationarity, multi-scale dynamics, and inter-variable dependencies make continuous multi-step forecasting challenging. This paper proposes MSTGraph-RAR for multivariate multi-step offshore wind power forecasting. Continuous series are segmented into patches, and the Multi-Scale Temporal–Variable Graph (MSTGraph) captures short- and long-range dynamics while learning scale-specific directed dependencies between power and meteorological variables. A Transformer encoder then generates base forecasts. Ramp-Aware Auxiliary Learning (RAL) uses future Down, Normal, and Up power-change states as auxiliary supervision to enhance state-sensitive representations, while Ramp-Aware Residual Correction (RRC) selectively refines forecasts through gated and bounded residual correction. Experiments on five real offshore wind farms show that MSTGraph-RAR achieves the best performance across all forecasting horizons and metrics. Compared with baseline models, MSE, MAE, and RMSE decrease by up to 36.13%, 24.80%, and 19.71%, respectively, while R^2 increases by up to 0.13. Ablation and ramp-intensive analyses further confirm the effectiveness of the proposed components and improved tracking of rapid power variations. Overall, MSTGraph-RAR provides an effective approach for reliable offshore wind power forecasting under complex non-stationary conditions.

Keywords: Offshore wind power forecasting; Graph neural networks; Multivariate time series forecasting; Multi-scale feature learning; Wind power ramp events

1. Introduction

As the scale of grid-integrated offshore wind power continues to expand, the variability and uncertainty of its power output increasingly affect power balancing, operational scheduling, and reserve capacity

© 2026 The authors. This is an open access article under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

allocation in power systems [1–5]. Accurate and reliable offshore wind power forecasting provides an essential foundation for operational decision-making and secure and economical power system operation [2,5]. Despite continued advances in forecasting methods, maintaining high forecasting accuracy across complex operating conditions and different forecasting horizons remains challenging [6].

This challenge is closely associated with the complex dynamics of offshore wind power time series. Continuous changes in marine meteorological conditions and turbine operating states give rise to pronounced non-stationarity, with power variation patterns differing considerably across operating conditions [3,6]. Meanwhile, slowly varying trends, periodic fluctuations, and rapid local variations intertwine across multiple temporal scales, and power output depends complexly on historical power and multiple meteorological variables [3,6,7]. Rapid power variations, particularly ramp events, are generally more difficult to forecast accurately [8–10]. Furthermore, uncertainty in future power variations and forecasting errors tend to increase as the forecasting horizon extends [11,12]. Together, these factors increase the modeling difficulty of multivariate multi-step offshore wind power forecasting.

As wind power forecasting research has advanced, forecasting methods have gradually evolved from conventional statistical approaches toward deep temporal representation learning. Methods such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and Kalman filtering offer transparent structures and high computational efficiency, but have limited representation capabilities under complex nonlinear and non-stationary conditions. Conventional machine learning methods, including support vector machines (SVMs) and random forests, improve nonlinear fitting but remain limited in learning hierarchical representations of complex temporal features [4,12]. Subsequently, deep learning models, including recurrent neural networks (RNNs), long short-term memory (LSTM) networks, convolutional neural networks (CNNs), temporal convolutional networks (TCNs), and Transformers, have progressively strengthened the modeling of both local and long-range temporal dependencies [4,12,13]. Building on these advances, FEDformer, iTransformer, and PatchTST further improve complex time-series forecasting through frequency-domain modeling, variable-wise representation, and patch-based representation, respectively [14–16]. Recent offshore studies have likewise combined LiDAR–SCADA fusion, multi-sensor inputs, dynamic time-delay estimation, modular multi-step design, and multi-scale attention mechanisms for short- and multi-step forecasting [17–21].

Beyond temporal dependencies, multivariate wind power forecasting also requires modeling the information interactions between historical power and meteorological variables. Accordingly, graph neural networks have increasingly been employed to learn relationships among variables, wind turbines, or wind farms, leading to adaptive, directed, and dynamic graph modeling approaches [22–24]. More specifically, MAGNN learns inter-variable relationships at different scales through a multi-scale temporal pyramid, whereas MSGNet combines frequency-domain scales with adaptive graph convolution to model scale-dependent inter-series relationships [25,26]. These studies demonstrate the effectiveness of jointly modeling multi-scale temporal information and variable relationships. However, existing graph-based wind power forecasting studies have primarily used graph structures to describe spatial relationships among turbines or wind farms. Research that jointly treats historical power and multiple meteorological factors as variable nodes and characterizes their interactions across different temporal receptive fields remains relatively limited.

Beyond temporal and inter-variable relationship modeling, existing studies have employed signal decomposition, similar-pattern matching, wind-regime clustering, and state-aware modeling to adapt to different power evolution patterns under non-stationary conditions [2,3,6,7,27,28]. For ramp events characterized by rapid changes, event classification, class-imbalance learning, and multi-task forecasting have been further adopted to model ramp occurrence and associated attributes, including magnitude, duration, and direction [8–10]. These studies indicate that ramp states contain important dynamic information distinct from normal power variations. However, existing work has primarily treated such information as a standalone

target for event identification or attribute prediction, with relatively limited attention to how ramp-state information can contribute to representation learning for continuous multi-step power forecasting. Therefore, incorporating ramp-related states into forecasting models as auxiliary information offers a potential means of learning more discriminative features for rapidly changing intervals [8–10].

As the forecasting horizon increases, forecasting errors generally become larger [11,12], while their characteristics vary across forecasting horizons, samples, and individual future steps. Existing residual-correction methods can further compensate for systematic biases in base forecasts [29]. However, applying uniform or unconstrained corrections to all forecasting positions may perturb predictions that are already accurate. Therefore, adaptively allocating the correction intensity while preserving reliable information in the base forecasts warrants further investigation in multi-step forecasting.

To address these challenges, this study develops MSTGraph-RAR for multivariate multi-step offshore wind power forecasting. MSTGraph jointly models temporal features and inter-variable relationships at the patch level across different receptive fields. Ramp-state auxiliary supervision enhances the shared representations, and gated, bounded residual correction refines the base forecasts. These modules operate at the representation, auxiliary-supervision, and output-correction stages, respectively. The main contributions are summarized as follows:

(1) To address the coexistence of multi-scale temporal dynamics and inter-variable dependencies in multivariate offshore wind power series, a Multi-Scale Temporal–Variable Graph learning module, termed MSTGraph, is developed. It extracts temporal context using fixed short- and long-range receptive fields over the same overlapping patch sequence and aggregates information from power and meteorological variables through separately learned sparse directed graphs, thereby enhancing cross-variable representations at the patch level.

(2) Ramp-Aware Auxiliary Learning (RAL) is introduced by formulating future power-change direction classification as an auxiliary task during training. This supervision enables the shared representations to better distinguish different power-change states, thereby enhancing the model's sensitivity to rapid variations such as ramp events and incorporating ramp-state information into representation learning for continuous multi-step power forecasting.

(3) To address the residual errors remaining in multi-step forecasting, a Ramp-Aware Residual Correction (RRC) module is designed. By combining historical power representations with the base forecasts, RRC adaptively determines the correction intensity for each future step and applies selective correction through bounded residuals, thereby reducing unnecessary perturbations to reliable base forecasts. The proposed framework is systematically evaluated across multiple forecasting horizons through ablation analyses, ramp-intensive scenarios, and cross-farm experiments, validating its forecasting performance, the effects of its individual modules, and its generalization capability.

Sections 2–5 present the methodology, experimental setup, results, and conclusions, respectively.

2. Methodology

2.1. Overall Architecture

The overall architecture of the proposed offshore wind power forecasting framework is illustrated in Figure 1. Given a historical multivariate time series, instance normalization is first applied independently to each variable. Following the patch-based representation strategy of PatchTST [16], overlapping patch segmentation, linear projection, and positional encoding are then performed to obtain patch-level latent representations:

$$H^{(0)} = \mathrm{PatchEmbed}\big(\mathrm{RevIN}(X)\big) \in \mathbb{R}^{B\times N\times P\times D} \tag{1}$$

where B, N, P, and D denote the batch size, number of variables, number of patches, and dimensionality of the latent representations, respectively.

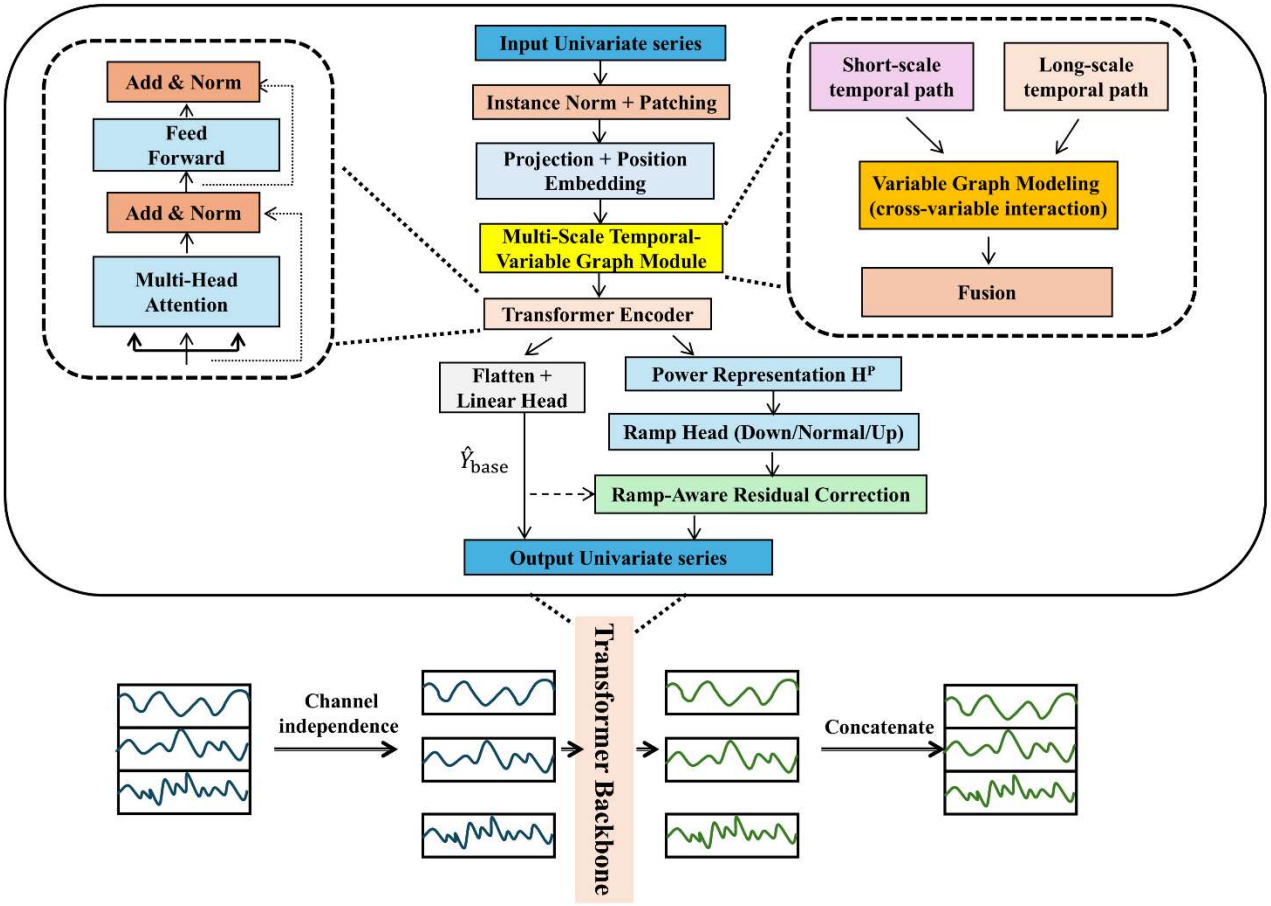


Figure 1. Overall architecture of the proposed offshore wind power forecasting framework. Within the Transformer encoder, solid arrows indicate the main flow of information, and dotted arrows denote residual connections.

The Multi-Scale Temporal–Variable Graph (MSTGraph) module is subsequently introduced to jointly model multi-scale temporal features and inter-variable dependencies through feature extraction at different temporal scales and variable-graph propagation. The enhanced representations are then fed into a channel-independent Transformer to capture long-range temporal dependencies within each variable, after which a forecasting head generates the base wind power forecasts. In addition to the primary forecasting task, Ramp-Aware Auxiliary Learning (RAL) and Ramp-Aware Residual Correction (RRC) are incorporated into the framework. RAL uses the latent representations of the Power channel to classify the Down, Normal, and Up states, thereby enhancing the model's ability to represent rapid power variations through auxiliary supervision. RRC then combines the latent Power features with the base forecasts to apply gated and bounded residual correction to the remaining forecasting errors, producing the final wind power forecasts.

2.2. Multi-Scale Temporal–Variable Graph Module

Offshore wind power time series exhibit variation patterns across multiple temporal scales and are jointly influenced by wind speed, wind direction, and other meteorological variables. However, a channel-independent Transformer primarily models temporal dependencies within each variable and lacks explicit inter-variable information exchange. To address this limitation, MSTGraph is inserted between the patch embedding layer and the Transformer encoder to enhance the initial patch representations through scale-specific temporal feature extraction and variable-graph propagation:

$$H = H^{(0)} \in \mathbb{R}^{B \times N \times P \times D} \quad (2)$$

As illustrated in Figure 2, MSTGraph employs two parallel branches along the patch dimension to capture temporal patterns over short- and long-range receptive fields. In each branch, layer normalization is first applied to the input, followed sequentially by depthwise convolution, pointwise convolution, GELU activation, and dropout. A residual connection is then used to preserve the original representation:

$$H_s^{temp} = H + \Phi_s\big(\mathrm{LN}(H)\big), s \in \text{short,long} \tag{3}$$

$$\Phi_s = \mathrm{Dropout}\Big(\mathrm{GELU}\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{DWConv}k_s, d_s(\cdot)\big)\Big)\Big) \tag{4}$$

The short-term branch uses a kernel size of $k = 3$ and a dilation rate of $d = 1$ to capture relatively local temporal context. The long-term branch adopts $k = 5$ and $d = 2$, yielding a receptive field that spans nine patch positions and captures temporal information over a broader range.

At each scale, MSTGraph treats the N input variables as graph nodes and constructs a graph score matrix using two independent sets of learnable node embeddings [30]:

$$L_s = \mathrm{ReLU}\big(E_{1,s}E_{2,s}^{\mathrm{T}}\big) \tag{5}$$

While retaining self-loops, the scores in each row are sparsified using Top-K selection [31], and a row-wise Softmax operation is subsequently applied to obtain the normalized adjacency matrix:

$$A_s = \mathrm{Softmax}_{row}\Big(\mathrm{TopK}_{self}(L_s)\Big) \tag{6}$$

Because $E_{1,s}$ and $E_{2,s}$ are learned independently, A_s is generally asymmetric and therefore represents a scale-specific learnable directed variable graph. The node embeddings are jointly optimized with the forecasting model but are independent of the input, resulting in a static learnable graph whose adjacency matrix is shared across samples and Patch positions after training. At each Patch position, the scale branch aggregates information along the variable dimension according to A_s:

$$H_s^{branch} = G_s\big(H_s^{temp}, A_s\big) \tag{7}$$

where the core adjacency aggregation can be expressed as $\sum_{j=1}^{N} A_{s,ij}\, H_{b,j,p,:}$. Therefore, temporal convolution and graph propagation operate along different dimensions: the former aggregates neighboring temporal information along the Patch axis within the same variable, whereas the latter exchanges information along the variable axis at a fixed Patch position.

Next, the representation increments of the two scale branches relative to the original input are computed and fused by arithmetic averaging:

$$\Delta H_s = H_s^{branch} - H \tag{8}$$

$$\Delta H_{fused} = \frac{\Delta H_{short} + \Delta H_{long}}{2} \tag{9}$$

The fused result is injected into the original representation through feature-wise gating:

$$H^g = H + \sigma(g) \odot \Delta H_{fused} \tag{10}$$

where $g \in \mathbb{R}^{1\times1\times D}$, the residual bypass retains the original Patch representation, while $\sigma(g)$ controls the injection strength of graph-enhanced information across different hidden feature dimensions. Thus, MSTGraph provides the subsequent Transformer with an enhanced representation that simultaneously contains multi-range temporal context and explicit cross-variable interactions.

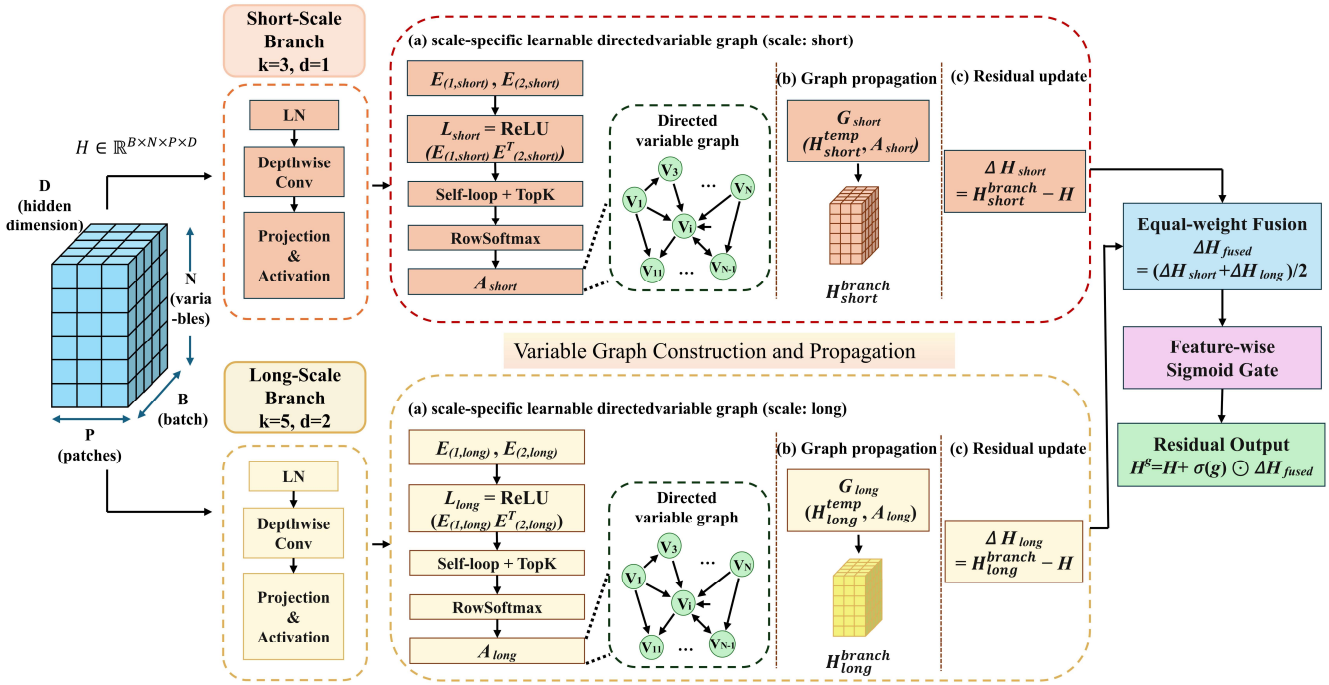


Figure 2. Architecture of the proposed Multi-Scale Temporal-Variable Graph (MSTGraph) module.

2.3. Ramp-Aware Auxiliary Learning

Offshore wind power series exhibit pronounced overall non-stationarity, but their local dynamics are clearly imbalanced: a large proportion of time windows show relatively smooth power evolution, whereas a small number of rapidly changing intervals, such as ramp events, show stronger non-stationarity and often constitute major sources of forecasting error.

To enhance the shared representation's ability to distinguish different power-change states, Ramp-Aware Auxiliary Learning (RAL) is introduced, with future power-change direction identification formulated as an auxiliary classification task. The overall architecture of RAL is illustrated in Figure 3a.

The latent representation of the Power channel output by the Transformer encoder is $H^P \in \mathbb{R}^{B\times P\times D}$. RAL extracts its global average representation and last-Patch representation and concatenates them into a power-state feature:

$$z^P = \mathrm{Concat}\left(\frac{1}{P}\sum_{p=1}^{P} H^P_{:,p,:}, H^P_{:,P,:}\right) \tag{11}$$

The ramp-aware classification head then outputs, for each forecast step, the probabilities of belonging to the Down, Normal, and Up states:

$$\pi = \mathrm{Softmax}\left(f_{RAL}(z^P)\right) \in \mathbb{R}^{B\times T\times 3} \tag{12}$$

where T denotes the forecasting length. To construct the state label for each forecast step, step-wise power changes are calculated using a threshold-based ramp-state definition [8–10]. The first change is relative to the last historical observation, while subsequent changes are computed between consecutive future values:

$$r_{b,t} = \begin{cases} Y_{b,1} - X^P_{b,L}, & t = 1, \\ Y_{b,t} - Y_{b,t-1}, & t = 2, \ldots, T. \end{cases} \qquad c_{b,t} = \begin{cases} \text{Down}, & r_{b,t} < \tau_{\mathrm{d}} \\ \text{Normal}, & \tau_{\mathrm{d}} \le r_{b,t} \le \tau_{\mathrm{u}} \\ \text{Up}, & r_{b,t} > \tau_{\mathrm{u}} \end{cases} \tag{13}$$

Here, $\tau_d = -\tau$ and $\tau_u = \tau$. The positive threshold τ is determined exclusively from the training set according to a prescribed quantile of the absolute power-change distribution. To mitigate class imbalance,

a class-weighted cross-entropy loss is introduced as auxiliary supervision. The MSTGraph-enhanced Transformer backbone, forecasting head, and RAL classification head are jointly optimized using the following objective:

$$\mathcal{L}_{stage1} = \mathcal{L}_{pred} + \lambda_{RAL}\mathcal{L}_{RAL}, \qquad \mathcal{L}_{RAL} = -\frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T} w_{c_{b,t}} \log \pi_{b,t,c_{b,t}}, \qquad w_c \propto n_c^{-1/2} \tag{14}$$

Here, $\mathcal{L}_{pred}$ denotes the mean squared error between the base forecasts $\hat{Y}_{base}$ and the observed power Y, while $\mathcal{L}_{RAL}$ denotes the class-weighted cross-entropy loss. The two losses jointly optimize the shared backbone, whereas the forecasting and classification heads are supervised by their respective objectives. The auxiliary loss coefficient is set to 0.05, and the class weights are calculated from the training-set frequencies using inverse-square-root weighting. The optimal backbone checkpoint is selected according to the validation forecasting MSE rather than the combined training loss. Through this auxiliary supervision, RAL encourages the shared representation to distinguish different power-change states, thereby complementing MSTGraph with transition-related information.

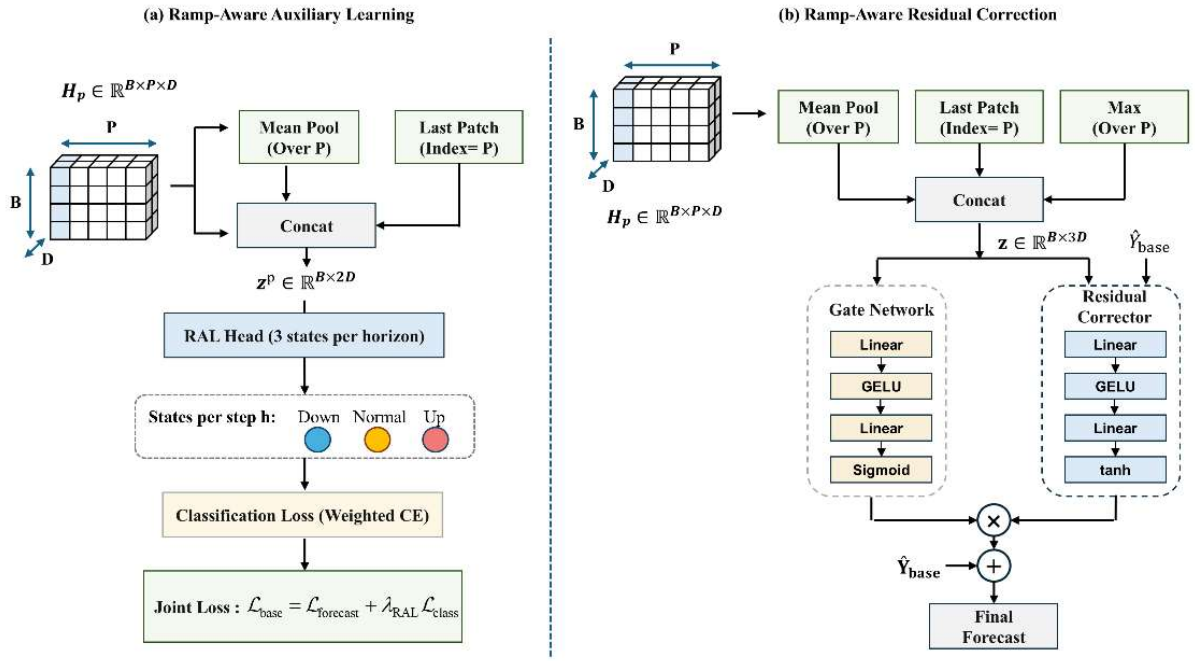


Figure 3. Architectures of (**a**) Ramp-Aware Auxiliary Learning (RAL) and (**b**) Ramp-Aware Residual Correction (RRC).

2.4. Ramp-Aware Residual Correction

As the forecasting horizon increases, residual forecasting errors remain, while unconstrained residual correction may improve abnormal-change regions at the cost of degrading a large number of Normal-state predictions. Therefore, this paper introduces Ramp-Aware Residual Correction (RRC), which applies selective and bounded output correction to reduce ramp-related errors while controlling the correction imposed on Normal regions. The overall architecture of RRC is illustrated in Figure 3b.

RRC jointly uses the Power representation H^P output by the Transformer and the base forecast $\hat{Y}_{base}$. First, Mean, Last, and Max pooling are performed on H^P along the Patch dimension and concatenated to obtain a historical state summary h_s. The gating network then generates a correction weight for each future forecast step:

$$p = \sigma\left(f_g(h_s)\right) \tag{15}$$

The residual corrector further combines the state summary with the base forecast and adopts a bounded output:

$$C = c_{max} \tanh\left(f_c([h_s, \hat{Y}_{base}])\right) \tag{16}$$

Here, c_{max} denotes the maximum residual correction magnitude, determined as the 95th percentile of the absolute base-forecast errors on the training set. The tanh activation bounds the predicted residual correction within $(-c_{max}, c_{max})$, preventing excessively large adjustments to the base forecasts.

The final forecast is given by

$$\hat{Y}_{final} = \hat{Y}_{base} + p \odot C \tag{17}$$

After backbone training, its parameters are frozen, and only the RRC module is optimized. The RRC training objective incorporates the final forecasting error, ramp-region residual learning, gate supervision, and a normal-region correction penalty:

$$\mathcal{L}_{RRC} = \mathcal{L}_{final} + \lambda_c \mathcal{L}_{corr} + \lambda_g \mathcal{L}_{gate} + \lambda_n \mathcal{L}_{normal} \tag{18}$$

Here, $\mathcal{L}_{final}$ denotes the MSE between the corrected forecasts and the ground truth. $\mathcal{L}_{corr}$ employs SmoothL1 loss to learn the base-forecast residuals at true ramp positions, while $\mathcal{L}_{gate}$ uses weighted binary cross-entropy to supervise the correction gate. $\mathcal{L}_{normal}$ suppresses unnecessary corrections during Normal states. The corresponding loss weights are set to 1, 1, 0.1, and 1, respectively. Unlike the first-stage objective, this loss updates only the parameters of RRC, leaving the trained forecasting backbone unchanged.

3. Data and Experimental Setup

To systematically evaluate the proposed model for multi-step offshore wind power forecasting, a unified experimental workflow is established, encompassing raw-data preprocessing, chronological data splitting, model training and validation, hyperparameter selection, and evaluation on an independent test set. The overall experimental workflow is illustrated in Figure 4. All models use identical data splits and evaluation protocols to ensure fair comparisons.

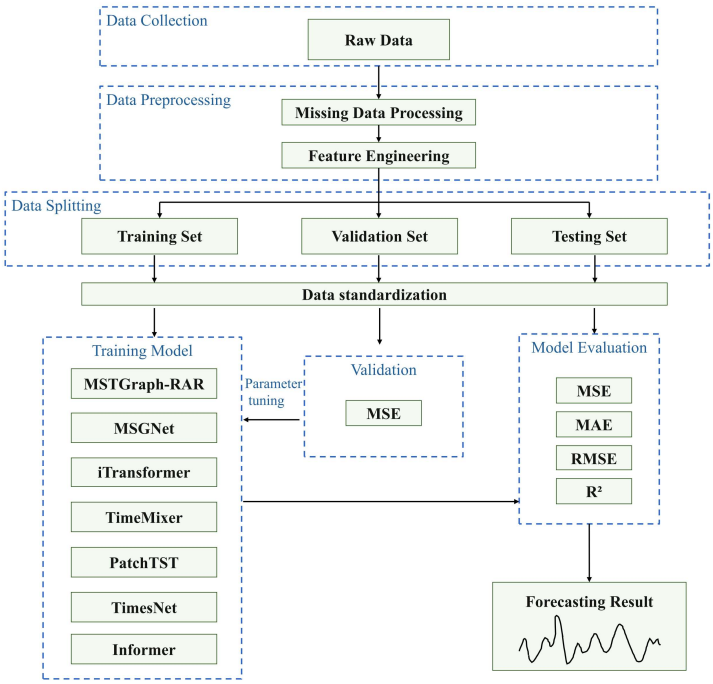


Figure 4. Overall experimental workflow for multi-step offshore wind power forecasting.

3.1. Dataset Description and Characteristics

This study uses the publicly available offshore wind power forecasting dataset released by the Digital China Innovation Contest. The raw dataset is available from the official competition website: https://www.dcic-china.com/competitions/10098/datasets (accessed on 21 May 2026). The dataset comprises observations from five real-world offshore wind farms in Fujian Province, China, hereafter denoted as S1–S5. It covers the period from 00:00 on 3 January 2022, to 23:45 on 30 April 2023, with a sampling interval of 15 min, yielding 231,840 observations in total. The input variables comprise ten meteorological factors, including atmospheric pressure, relative humidity, cloud cover, wind speeds and directions at heights of 10 and 100 m, temperature, solar irradiance, and precipitation. The prediction target is the active power output time series of each wind farm.

From a temporal perspective, wind power exhibits strong short-term continuity, whereas its temporal correlation decays rapidly as the lag increases. Across the five wind farms, the mean autocorrelation coefficients of Power are 0.991 and 0.954 at lags of 15 min and 1 h, respectively, but decrease to 0.742 and 0.454 at lags of 6 h and 24 h. The seven-day trajectories in Figure 5a further reveal alternating patterns of smooth variations, abrupt transitions, and sustained high- or low-power operating regimes. These observations indicate that local temporal continuity alone is insufficient for forecasting over extended horizons.

At the cross-farm level, substantial heterogeneity is observed in the power-output scales of the five wind farms. As shown in Figure 5b, the minimum and maximum mean power outputs are 10.571 MW at S3 and 115.205 MW at S2, respectively, while the maximum power output ranges from 44.43 to 296.58 MW across the farms. Pairwise correlation analysis of the power time series yields Pearson correlation coefficients ranging from 0.453 to 0.786. These results suggest that power evolution across the wind farms is driven by common regional meteorological processes while retaining pronounced farm-specific response characteristics. Accordingly, for forecasting this multi-farm panel dataset, the power sequences should not be directly concatenated and treated as a single univariate series. Instead, the model should learn both shared temporal patterns across farms and farm-specific local characteristics associated with different power-output scales.

At the variable level, active power exhibits nonlinear dependencies on multiple meteorological variables. WindSpeed10m and WindSpeed100m show the strongest linear associations with Power, with Pearson correlation coefficients of 0.581 and 0.572, respectively. Their Spearman correlation coefficients increase to 0.779 and 0.790, indicating that the monotonic relationships between wind speed and power are considerably stronger than their simple linear associations. Variables such as atmospheric pressure, wind direction, temperature, and cloud cover exhibit relatively weak univariate linear correlations with Power. The correlation structure shown in Figure 5c suggests that processing each channel independently is insufficient to fully characterize the power-generation process, underscoring the need to explicitly model inter-variable interactions.

From an operational perspective, large changes in offshore wind power are relatively infrequent but are often challenging to forecast. To characterize different power-change states, the power observations are first standardized using a common scaler fitted to the pooled training data from all five wind farms. Adjacent 15-min power increments are then calculated within continuous training segments of each farm and pooled to form a common training-set power-change distribution. The 90th percentile of the pooled absolute power increments (Pooled Q90) is adopted as the shared ramp threshold, yielding $\tau = 0.089794$ in standardized units, equivalent to a power change of 5.44 MW over a 15-min interval. Under this threshold, the Down, Normal, and Up states comprise 8130, 146,056, and 8094 valid training increments, accounting for 5.01%, 90.00%, and 4.99%, respectively. Figure 5d presents the cumulative distribution of the absolute power increments.

To further characterize the farm-level state distributions under the shared threshold, Table 1 reports the proportions of Down, Normal, and Up states for each farm. These statistics are calculated exclusively from consecutive power increments within continuous training segments before sliding-window construction, with each valid increment counted once. Each farm contributes 32,456 valid increments, resulting in 162,280 increments in total. The ramp-state proportions differ substantially across farms: the combined proportion of Down and Up states ranges from 0.88% for S5 to 38.04% for S2. These differences show that a shared absolute threshold yields different ramp-state proportions across farms with distinct power-change scales and increment distributions. Section 4.4 further evaluates the effect of threshold selection on forecasting performance through sensitivity experiments comparing pooled and farm-specific thresholds.

Taken together, Figure 5 characterizes the dataset from four complementary perspectives: temporal variations, cross-farm heterogeneity, inter-variable relationships, and the sparsity of ramp events. These characteristics highlight the major challenges of offshore wind power forecasting and provide an empirical basis for the architectural design of the proposed model.

Table 1. Training-set ramp-state distributions under the pooled Q90 threshold.

Farm	Down, *n* (%)	Normal, *n* (%)	Up, *n* (%)	Ramp (%)
S1	257 (0.79)	31,937 (98.40)	262 (0.81)	1.60
S2	6283 (19.36)	20,109 (61.96)	6064 (18.68)	38.04
S3	276 (0.85)	31,833 (98.08)	347 (1.07)	1.92
S4	1169 (3.60)	30,006 (92.45)	1281 (3.95)	7.55
S5	145 (0.45)	32,171 (99.12)	140 (0.43)	0.88
Overall	8130 (5.01)	146,056 (90.00)	8094 (4.99)	10.00

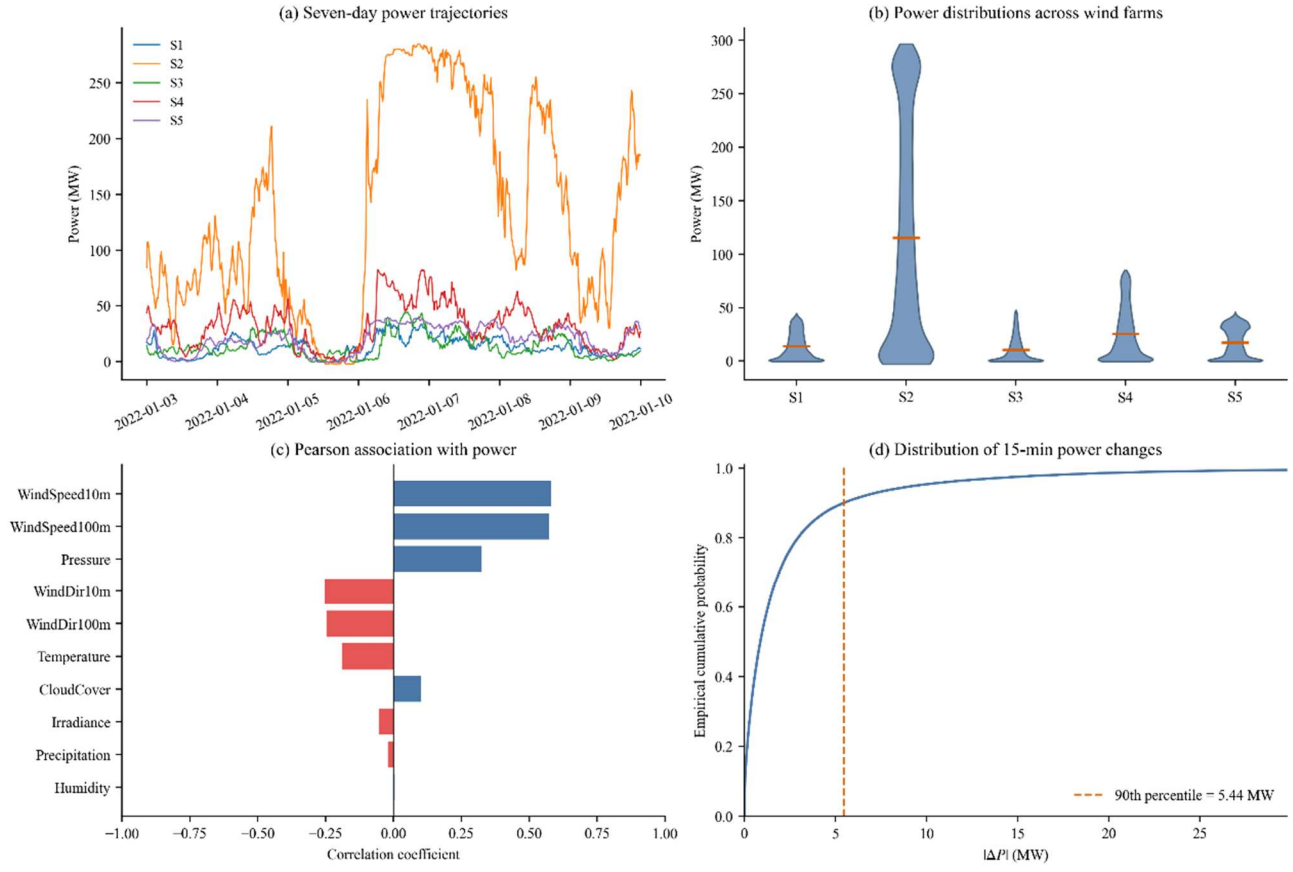


Figure 5. Characteristics of the five-farm offshore wind power panel dataset: (**a**) seven-day Power trajectories; (**b**) Power distributions across the five wind farms; (**c**) Pearson correlations between meteorological variables and Power; and (**d**) empirical cumulative distribution of absolute 15-min power increments in the training set.

3.2. Data Preprocessing

To ensure temporal consistency among observations from different offshore wind farms and prevent future-information leakage during preprocessing, the raw data undergo a unified procedure comprising data cleaning, chronological splitting, standardization, and supervised sample construction. First, the observations are chronologically ordered by wind farm and contiguous data segment, followed by checks for duplicate timestamps, non-finite values, and sampling continuity. The raw Power series contains 15 missing observations (0.00647%), all within the training subsets. After chronological data splitting, linear interpolation is applied only to short internal gaps within each wind farm and training subset, without crossing contiguous-segment or subset boundaries. The validation and test subsets contain no missing Power observations and undergo no interpolation, thereby avoiding future-information leakage through test-data imputation. After preprocessing, the ten meteorological variables and historical active power are jointly organized into an 11-dimensional multivariate input sequence, with active power serving as the prediction target.

To preserve temporal causality under realistic forecasting conditions, random splitting is not used. Instead, the data from each wind farm are chronologically divided into training, validation, and test sets at approximate ratios of 70%, 10%, and 20%, containing 162,285, 23,190, and 46,365 observations, respectively. The training set is used to learn the model parameters and estimate all training-dependent statistics, the validation set is used for early stopping and optimal model selection, and the test set is used exclusively for final performance evaluation after the model architecture and hyperparameters have been determined.

Given the substantial differences in units and numerical ranges between active power and the meteorological variables, standardization is applied to place all input features on a comparable scale. For variable j, the standardized value is defined as follows:

$$\tilde{x}_{t,j} = \frac{x_{t,j} - \mu_j^{train}}{\sigma_j^{train}} \tag{19}$$

where μ_j^{train} and σ_j^{train} denote the mean and standard deviation of variable j calculated from the training set, respectively. All standardization statistics are computed exclusively from the training data and then held fixed. The same parameters are applied to the validation and test sets, thereby preventing their distributional information from leaking into the model-training process. For analyses requiring physical interpretation, the predicted and observed Power values are inverse-transformed to the original MW scale using the training-set statistics.

3.3. Experimental Setup

The input sequence length is fixed at 192 time steps, corresponding to 48 h of historical observations. Four forecasting horizons of 12, 24, 48, and 96 steps are considered, corresponding to 3, 6, 12, and 24 h, respectively. The patch length and stride are set to 16 and 8, resulting in 24 patches for each input sequence. The model has a hidden dimension of 128, with eight attention heads, two encoder layers, and a feed-forward network dimension of 256. The dropout rate is set to 0.1. All base models are trained with the Adam optimizer, an initial learning rate of 0.001, and a OneCycle learning-rate schedule.

For MSTGraph, the node-embedding dimension is 16, and $K = 5$ connections are retained per node, including one self-loop and four non-self neighbors. The ramp threshold is the 90th percentile of absolute training-set power increments computed within each continuous segment. For the five-farm offshore wind power panel dataset split, this threshold is 0.089794 in standardized units, equivalent to 5.44 MW per 15-min interval. The correction bound $c_{\max}$ is the 95th percentile of absolute base-forecast residuals on the training set and is determined separately for each fitted base model.

Computational requirements arise from temporal encoding, variable-graph propagation, and the two-stage training procedure. The experiments use 24 patch tokens derived from 192-step inputs and a batch size of 64 on an NVIDIA GeForce RTX 4070 Laptop GPU with FP32 precision.

The main experimental configurations are summarized in Table 2.

Table 2. Main experimental configurations.

Item	Configuration
Dataset	Offshore wind power panel dataset containing 231,840 samples from five wind farms
Sampling interval	15 min
Input features	Ten meteorological variables and historical power time series
Forecasting task	Future wind power time-series forecasting
Data split	Training/validation/test sets: 162,285/23,190/46,365 samples (70%/10%/20%)
Input sequence length	192 time steps (48 h)
Forecasting horizon	12, 24, 48, and 96 time steps (3, 6, 12, and 24 h)
PatchTST hyperparameters	Patch length: 16; stride: 8; number of patches: 24; hidden dimension: 128; attention heads: 8; encoder layers: 2; feed-forward dimension: 256
Optimizer	Adam; initial learning rate: 0.001; OneCycle learning-rate scheduler; batch size: 64
Training strategy	early-stopping patience: 3
Repeated experiments	Random seeds: 2024, 2025, and 2026
Hardware and software	NVIDIA GeForce RTX 4070 Laptop GPU; Python 3.8.20; PyTorch 1.11.0; CUDA 11.3; FP32 precision

3.4. Evaluation Metrics

Forecasting performance is evaluated using mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). MSE and RMSE assign greater weight to large forecasting errors and therefore reflect the model's ability to limit substantial deviations. MAE measures the average absolute deviation between the predicted and actual power values, while R^2 evaluates the model's goodness of fit to variations in wind power. Lower MSE, MAE, and RMSE values indicate higher forecasting accuracy, whereas an R^2 value closer to 1 indicates stronger explanatory and fitting performance. These metrics are defined as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{20}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{21}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{22}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{23}$$

where y_i and $\hat{y}_i$ denote the actual and predicted values, respectively; n is the number of samples; and y represents the mean of the actual values, calculated as $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Overall metrics use micro-averaging over these pairs pooled across the five wind farms.

4. Experimental Results and Discussion

4.1. Overall Forecasting Performance

To comprehensively evaluate the effectiveness of the proposed model for offshore wind power forecasting, comparative experiments are conducted against several representative baseline models across multiple forecasting horizons. The selected baselines cover several major categories of time-series forecasting methods, including Transformer-based models (iTransformer [15], PatchTST [16], and Informer [32]), multi-scale temporal modeling approaches (TimesNet [33] and TimeMixer [34]), and a graph-enhanced forecasting model (MSGNet [26]). Furthermore, considering the strong short-term autocorrelation of wind power, persistence forecasting, which uses the last observed power value to predict every step within the forecasting horizon, is also included as a comparative method. Four forecasting horizons of 12, 24, 48, and 96 steps are considered to evaluate model performance in different forecasting horizons.

As shown in Table 3 and Figure 6, MSTGraph-RAR consistently achieves the best performance across all forecasting horizons and evaluation metrics. Compared with the baseline models, the proposed method delivers consistent error reductions for the 12-, 24-, 48-, and 96-step forecasting tasks, with maximum reductions of 36.13% in MSE, 24.80% in MAE, and 19.71% in RMSE, together with an absolute increase of up to 0.13 in R^2.

These results demonstrate that MSTGraph-RAR effectively enhances the representation of dynamic variations in offshore wind power and maintains reliable forecasting performance over extended horizons.

Figures 7–11 present the power forecasting trajectories of MSTGraph-RAR and six representative baseline models over continuous test intervals from the five offshore wind farms. All methods are evaluated using identical test timestamps and ground-truth power sequences. Overall, differences among the models are most evident during rapid ramps, local peaks and troughs, sustained power plateaus, and abrupt reversals. Compared with the baseline models, MSTGraph-RAR maintains closer phase alignment and more consistent variation directions with the observed trajectories across most of the displayed intervals. It also exhibits less amplitude attenuation and response lag in several highly fluctuating regions. These results indicate that the model effectively tracks local non-stationary variations while preserving its forecasting capability during relatively stable periods.

Table 3. Forecasting errors across different forecasting horizons.

Metric	Steps	MSTGraph-RAR	MSGNet	iTrans-Former	TimeMixer	PatchTST	TimesNet	Informer	Persistence
MSE	12	0.076 ± 0.002	0.113 ± 0.006	0.099 ± 0.003	0.112 ± 0.001	0.084 ± 0.001	0.119 ± 0.002	0.091 ± 0.002	0.082 ± 0.000
	24	0.142 ± 0.002	0.182 ± 0.006	0.173 ± 0.007	0.158 ± 0.002	0.161 ± 0.002	0.206 ± 0.008	0.207 ± 0.020	0.157 ± 0.000
	48	0.250 ± 0.004	0.294 ± 0.007	0.296 ± 0.010	0.268 ± 0.007	0.279 ± 0.004	0.304 ± 0.007	0.363 ± 0.008	0.291 ± 0.000
	96	0.400 ± 0.003	0.456 ± 0.022	0.463 ± 0.006	0.425 ± 0.005	0.439 ± 0.003	0.496 ± 0.026	0.526 ± 0.006	0.502 ± 0.000
MAE	12	0.137 ± 0.001	0.169 ± 0.004	0.152 ± 0.002	0.175 ± 0.002	0.138 ± 0.002	0.172 ± 0.003	0.156 ± 0.006	0.142 ± 0.000
	24	0.185 ± 0.003	0.217 ± 0.003	0.207 ± 0.005	0.194 ± 0.004	0.199 ± 0.003	0.233 ± 0.006	0.246 ± 0.004	0.189 ± 0.000
	48	0.253 ± 0.003	0.282 ± 0.002	0.277 ± 0.007	0.260 ± 0.002	0.267 ± 0.002	0.280 ± 0.002	0.308 ± 0.006	0.255 ± 0.000
	96	0.332 ± 0.001	0.358 ± 0.004	0.357 ± 0.004	0.343 ± 0.005	0.351 ± 0.003	0.375 ± 0.006	0.394 ± 0.015	0.341 ± 0.000
RMSE	12	0.277 ± 0.004	0.337 ± 0.008	0.315 ± 0.004	0.335 ± 0.002	0.290 ± 0.002	0.345 ± 0.004	0.302 ± 0.004	0.286 ± 0.000
	24	0.376 ± 0.002	0.427 ± 0.007	0.416 ± 0.008	0.397 ± 0.003	0.401 ± 0.003	0.454 ± 0.009	0.455 ± 0.023	0.396 ± 0.000
	48	0.500 ± 0.004	0.542 ± 0.006	0.544 ± 0.009	0.518 ± 0.006	0.528 ± 0.003	0.551 ± 0.006	0.603 ± 0.007	0.539 ± 0.000
	96	0.632 ± 0.002	0.675 ± 0.016	0.680 ± 0.004	0.652 ± 0.004	0.662 ± 0.002	0.704 ± 0.019	0.725 ± 0.004	0.708 ± 0.000
R^2	12	0.922 ± 0.002	0.884 ± 0.006	0.898 ± 0.003	0.885 ± 0.001	0.914 ± 0.001	0.878 ± 0.002	0.906 ± 0.002	0.916 ± 0.000
	24	0.854 ± 0.002	0.813 ± 0.006	0.822 ± 0.007	0.838 ± 0.002	0.834 ± 0.002	0.788 ± 0.008	0.787 ± 0.021	0.839 ± 0.000
	48	0.743 ± 0.004	0.698 ± 0.007	0.696 ± 0.010	0.725 ± 0.007	0.713 ± 0.004	0.688 ± 0.007	0.627 ± 0.008	0.701 ± 0.000
	96	0.589 ± 0.003	0.532 ± 0.023	0.525 ± 0.006	0.564 ± 0.005	0.549 ± 0.003	0.491 ± 0.027	0.460 ± 0.006	0.485 ± 0.000

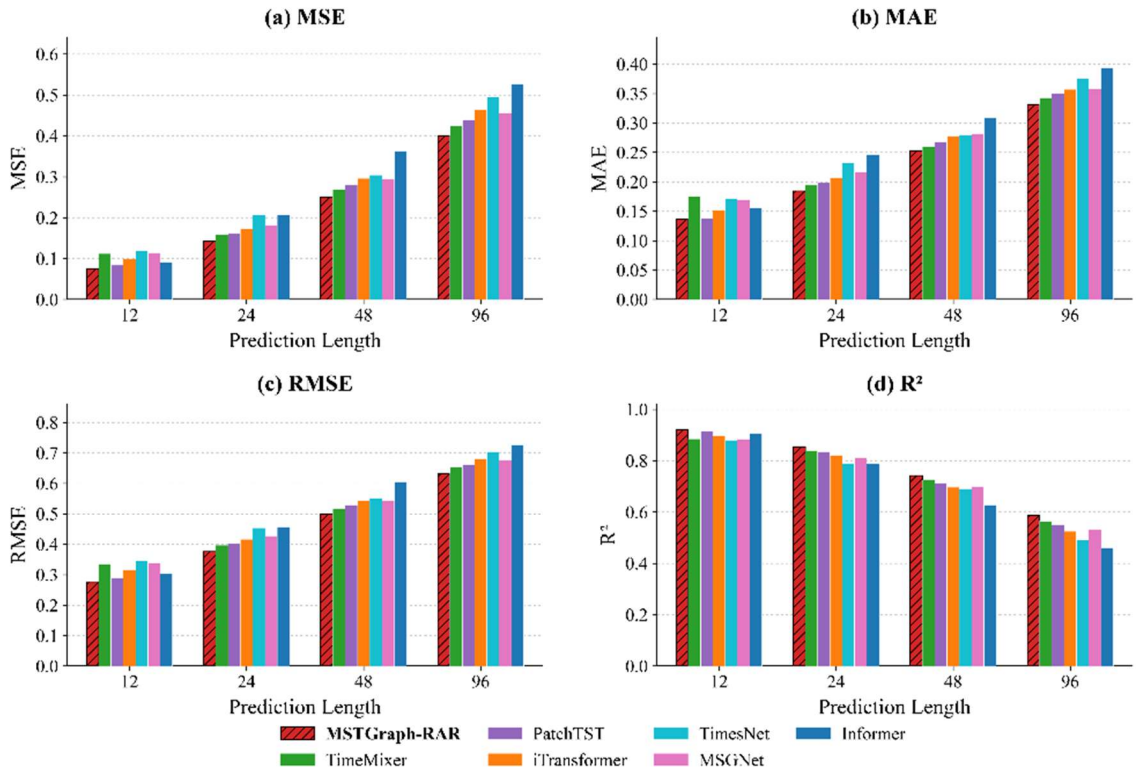


Figure 6. Forecasting performance of different models across forecasting horizons: (**a**) mean squared error (MSE); (**b**) mean absolute error (MAE); (**c**) root mean squared error (RMSE); and (**d**) coefficient of determination (R^2).

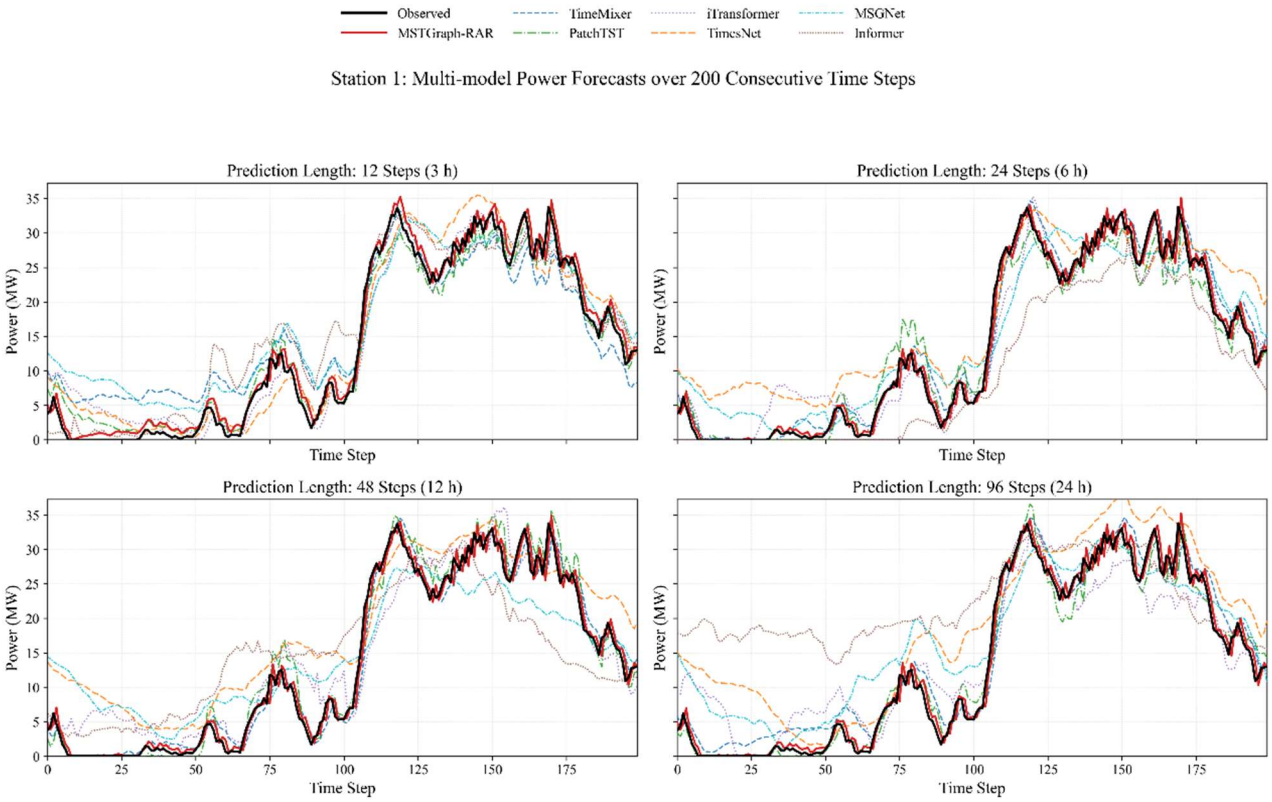


Figure 7. Multi-model power forecasting results for wind farm S1 under different forecasting horizons.

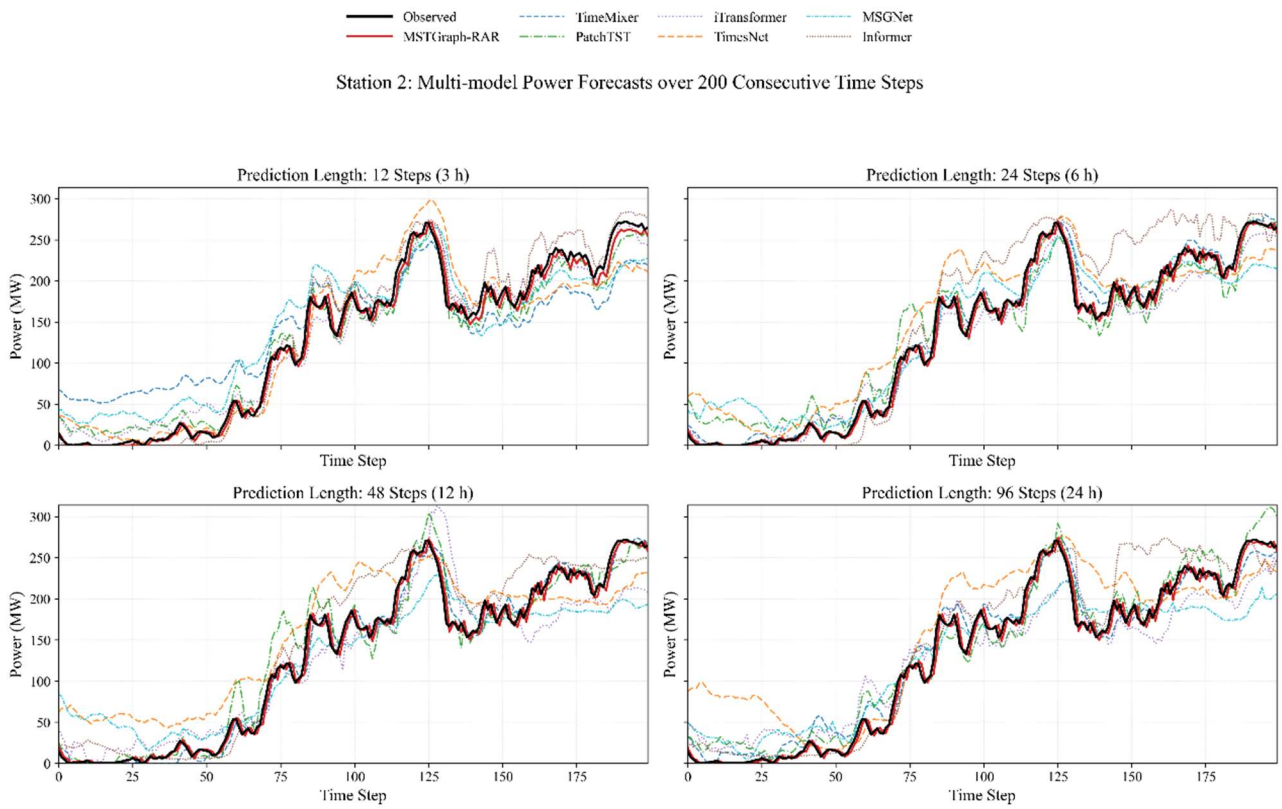


Figure 8. Multi-model power forecasting results for wind farm S2 under different forecasting horizons.

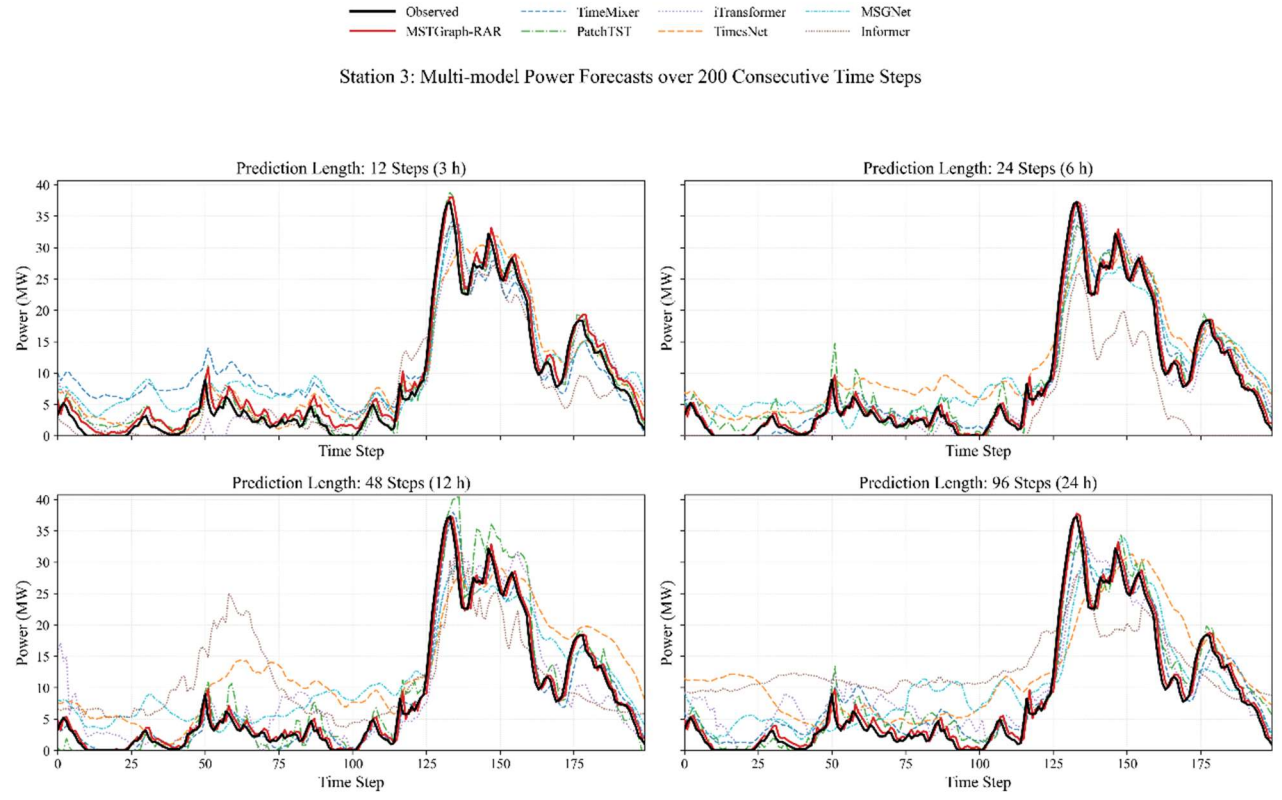


Figure 9. Multi-model power forecasting results for wind farm S3 under different forecasting horizons.

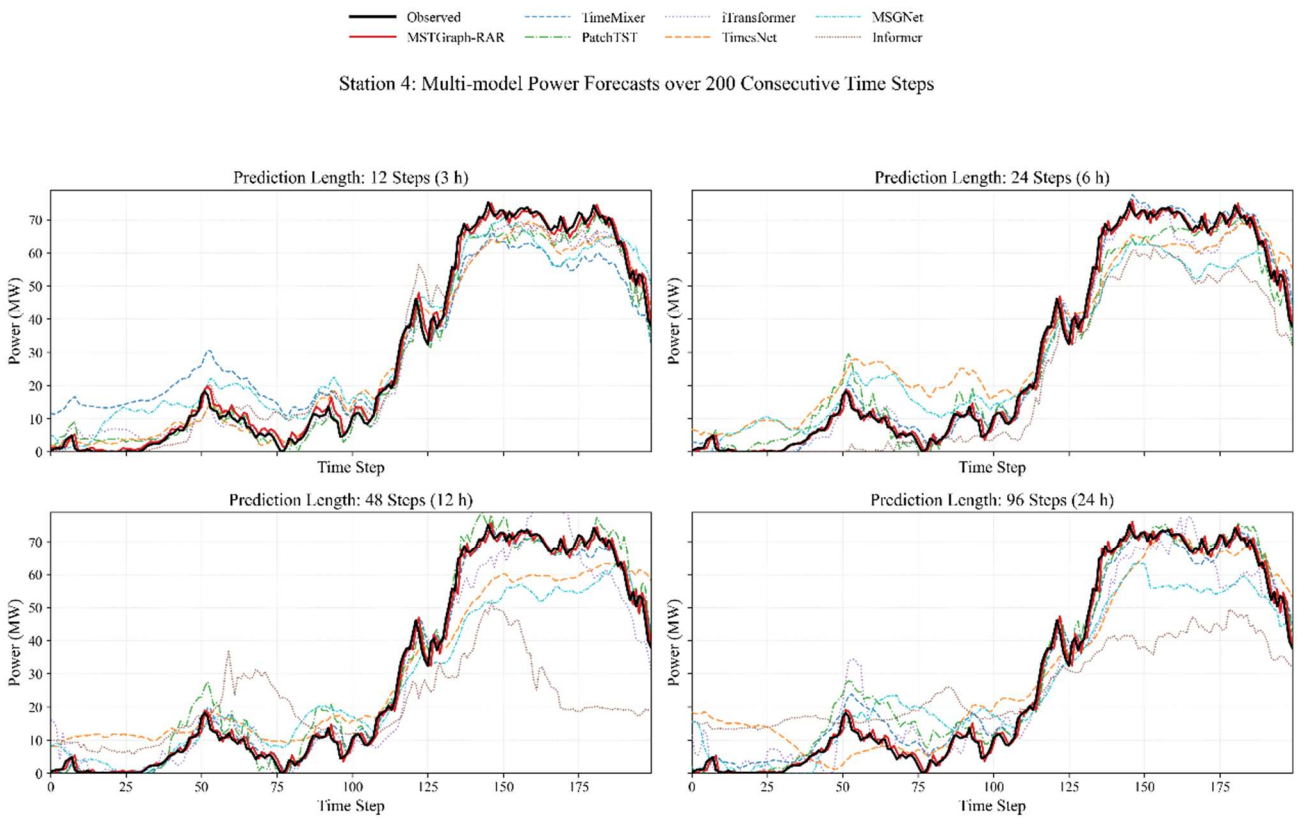


Figure 10. Multi-model power forecasting results for wind farm S4 under different forecasting horizons.

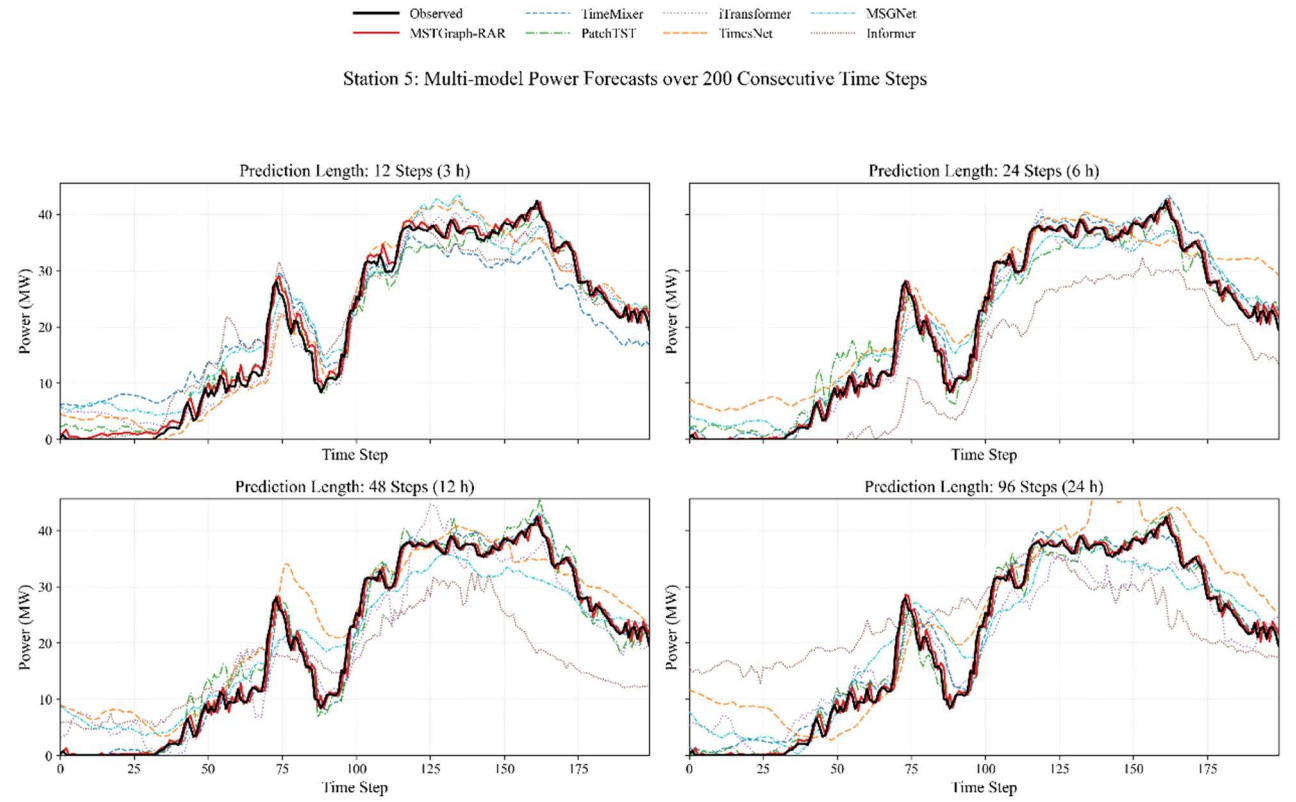


Figure 11. Multi-model power forecasting results for wind farm S5 under different forecasting horizons.

As shown in Figure 12, the relative-error distributions provide further insight into the forecasting behavior of the compared models. The central interval is defined as $-0.05 < e_{\mathrm{rel}} \leq 0.05$, whereas the two

extreme tails correspond to $e_{\mathrm{rel}} \leq -0.35$ and $e_{\mathrm{rel}} > 0.35$. Across the evaluated horizons, MSTGraph-RAR generally places more probability mass within the central interval than the baseline models. Its advantage is also evident in the extreme tails, where large forecasting deviations occur less frequently. Although the error distributions of all models broaden as the forecasting horizon increases, MSTGraph-RAR maintains a comparatively greater concentration around zero and lower tail frequencies. Overall, MSTGraph-RAR exhibits a more favorable relative-error distribution across the evaluated forecasting horizons.

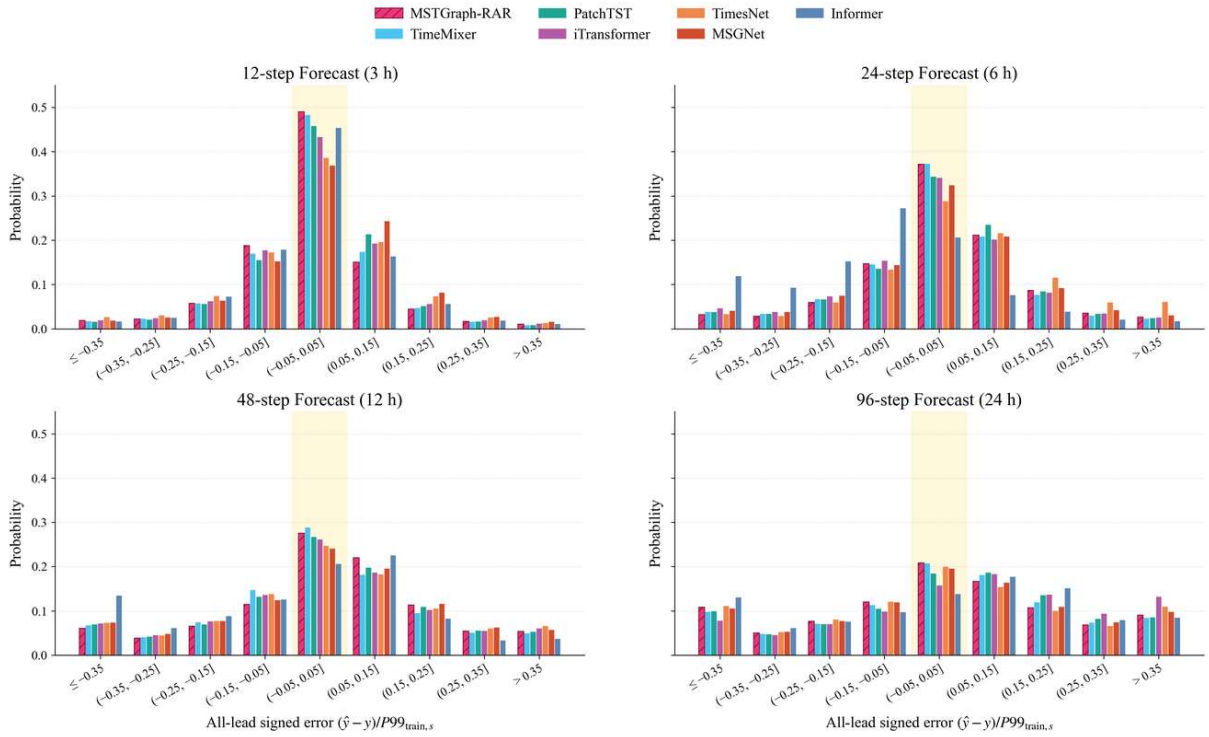


Figure 12. Histograms of the relative errors of different models across forecasting horizons.

4.2. Ablation Study

4.2.1. Quantitative Ablation Results

To quantitatively evaluate the contribution of each component to forecasting performance, ablation experiments are conducted at two levels. Module-level ablations examine the effects of MSTGraph, RAL, and RRC on overall forecasting performance, while component-level comparisons further assess the individual design choices within these modules.

For the module-level experiments, three ablation variants are constructed by removing the Multi-Scale Temporal–Variable Graph module (*w/o* MSTGraph), Ramp-Aware Auxiliary Learning (*w/o* RAL), and Ramp-Aware Residual Correction (*w/o* RRC), respectively. The PatchTST backbone is included as the reference model. All variants use identical data splits, input lengths, optimization settings, and evaluation procedures. Table 4 reports the mean and standard deviation obtained from three random seeds.

As shown in Table 4, the complete model achieves the lowest MSE at all four forecasting horizons. Compared with the PatchTST backbone, its MSE is reduced by 9.52%, 11.80%, 10.39%, and 8.88% for the 12-, 24-, 48-, and 96-step forecasting tasks, respectively, demonstrating consistent performance gains across different horizons. Among the three ablation variants, removing MSTGraph causes the largest performance degradation, increasing MSE by 7.89%, 9.86%, 8.40%, and 6.25% at the four respective horizons. This finding indicates that jointly modeling multi-scale temporal patterns and inter-variable dependencies is the primary source of the overall performance improvement. Removing RAL and RRC

increases MSE by 1.41–5.26% and 1.41–6.58%, respectively, with the average error increasing across all forecasting horizons.

Table 4. MSE comparison across different model variants.

Model Configuration	12	24	48	96
Complete model	0.076 ± 0.002	0.142 ± 0.002	0.250 ± 0.004	0.400 ± 0.003
w/o MSTGraph	0.082 ± 0.001	0.156 ± 0.003	0.271 ± 0.002	0.425 ± 0.007
w/o RAL	0.080 ± 0.002	0.144 ± 0.002	0.255 ± 0.006	0.415 ± 0.014
w/o RRC	0.081 ± 0.002	0.144 ± 0.002	0.259 ± 0.011	0.423 ± 0.007
PatchTST (baseline)	0.084 ± 0.001	0.161 ± 0.002	0.279 ± 0.004	0.439 ± 0.003

The internal contributions of MSTGraph are further examined by separately removing variable-graph propagation and multi-scale temporal mixing. The corresponding results are reported in Table 5.

Table 5. MSE comparison of MSTGraph component-level ablations across different forecasting horizons.

Model Configuration	12	24	48	96
Full MSTGraph	0.076 ± 0.002	0.142 ± 0.002	0.250 ± 0.004	0.400 ± 0.003
w/o Variable Graph	0.083 ± 0.001	0.157 ± 0.003	0.276 ± 0.002	0.421 ± 0.006
w/o Temporal Mixing	0.078 ± 0.002	0.144 ± 0.002	0.259 ± 0.006	0.409 ± 0.012

As shown in Table 5, removing variable-graph propagation increases the MSE by 9.21%, 10.56%, 10.40%, and 5.25% at the 12-, 24-, 48-, and 96-step forecasting horizons, respectively. In comparison, removing temporal mixing results in smaller MSE increases of 2.63%, 1.41%, 3.60%, and 2.25% at the corresponding horizons. These results indicate that variable-graph propagation makes a greater contribution to forecasting accuracy under the evaluated settings, while multi-scale temporal modeling provides additional improvements.

To further investigate the effectiveness of the proposed ramp-aware mechanisms, additional comparative experiments are conducted across four forecasting horizons. For RAL, the class-weighted variant is compared with an unweighted variant and a model without auxiliary classification. For RRC, the full module is compared with a model without residual correction, a conventional residual head, an ungated correction variant, and a variant without magnitude bounding. These comparisons assess the contributions of class weighting, residual correction, gating, and magnitude bounding to forecasting performance.

As shown in Table 6, RAL-Weighted achieves the lowest mean MSE across all four forecasting horizons. Compared with RAL-Unweighted, the weighted variant reduces MSE by 2.59%, 0.56%, 4.89%, and 5.65% for the 12-, 24-, 48-, and 96-step forecasting tasks, respectively. It also outperforms the variant without RAL at all evaluated horizons. These results indicate that class-weighted auxiliary supervision contributes to forecasting accuracy under the evaluated settings, although the magnitude of improvement varies across forecasting horizons.

As shown in Table 7, Full RRC reduces MSE by 5.80%, 0.97%, 3.51%, and 5.46% relative to the variant without RRC at the 12-, 24-, 48-, and 96-step forecasting horizons, respectively. It also achieves lower mean MSE than the conventional residual head and the variant without magnitude bounding across all four horizons, supporting the effectiveness of the complete correction design and bounded residual updates. Gating further improves mean accuracy at 12 and 24 steps and reduces variability across runs at 48 and 96 steps while maintaining comparable mean MSE. At 96 steps, for example, Full RRC and the ungated variant both achieve a mean MSE of 0.3998, while gating reduces the standard deviation from 0.0086 to 0.0027. Together, these results highlight the complementary roles of magnitude bounding and gating in improving forecasting accuracy and consistency across repeated runs.

Table 6. MSE comparison of RAL variants across different forecasting horizons.

Model Configuration	12	24	48	96
w/o RAL	0.0805 ± 0.002	0.1441 ± 0.0023	0.2555 ± 0.0064	0.4151 ± 0.0141
RAL-Unweighted	0.0811 ± 0.0018	0.1438 ± 0.0016	0.2640 ± 0.0115	0.4251 ± 0.0343
RAL-Weighted	0.0790 ± 0.0027	0.1430 ± 0.0004	0.2511 ± 0.0114	0.4011 ± 0.0068

Table 7. MSE comparison of RRC variants across different forecasting horizons.

Model Configuration	12	24	48	96
w/o RRC	0.0811 ± 0.0018	0.1438 ± 0.0016	0.2589 ± 0.0114	0.4229 ± 0.0068
Conventional Residual Head	0.0769 ± 0.0016	0.1432 ± 0.0012	0.2512 ± 0.0098	0.4012± 0.0010
RRC *w/o* Gating	0.0782 ± 0.0021	0.1438 ± 0.0014	0.2488 ± 0.0068	0.3998 ± 0.0086
RRC *w/o* Bounding	0.0800 ± 0.0024	0.1438 ± 0.0016	0.2508 ± 0.0044	0.4011 ± 0.0060
Full RRC	0.0764 ± 0.0024	0.1424 ± 0.0016	0.2498 ± 0.0040	0.3998 ± 0.0027

4.2.2. Analysis of Module Effects

To investigate the scale-specific inter-variable dependencies learned by MSTGraph, Figure 13 visualizes the directed adjacency matrices of the short- and long-scale branches, together with their element-wise absolute differences. Here, each adjacency matrix element denotes the aggregation weight assigned to information transmitted from a source variable to a receiving variable. At the short scale, the Power node primarily aggregates information from wind speed at 100 m (WS100), wind speed at 10 m (WS10), and atmospheric pressure (Pres), suggesting a greater emphasis on wind-related information within the shorter temporal receptive field. At the long scale, higher weights are assigned to solar irradiance (Irr), air temperature (Temp), and WS100, indicating that the longer-scale branch incorporates a broader range of meteorological information.

To further quantify the differences between the two scale-specific graphs, the Pearson correlation coefficient of their adjacency weights and the overlap score of their Top-K edge sets were calculated, yielding values of 0.490 and 0.594, respectively. These results indicate that the two branches capture partially shared variable dependencies while exhibiting distinct information-aggregation patterns at different temporal scales. Together with the ablation results, this analysis suggests that MSTGraph enhances the channel-independent representations learned by PatchTST through scale-specific inter-variable information exchange, contributing to improved forecasting performance.

To further examine whether the learned variable dependencies are consistently recovered under different training configurations, the adjacency matrices were compared across forecasting horizons and training-farm combinations. Across forecasting horizons, the mean adjacency-weight Pearson correlations are 0.714 and 0.612 for the short- and long-scale branches, respectively, with corresponding edge-set Jaccard similarities of 0.574 and 0.629. Across training-farm combinations, the corresponding Pearson correlations are 0.706 and 0.803, while the Jaccard similarities are 0.560 and 0.757. In addition, several connections to the Power node are repeatedly identified across different configurations. In the short-scale branch, Pres→Power and WS100→Power are retained with positive graph scores in 18 and 19 of the 20 examined models, respectively. In the long-scale branch, WS100→Power and Irr→Power appear in 20 and 19 models, respectively. These results suggest that certain inter-variable aggregation patterns remain consistent across forecasting horizons and training-farm combinations, further supporting the reproducibility of the learned graph structures.

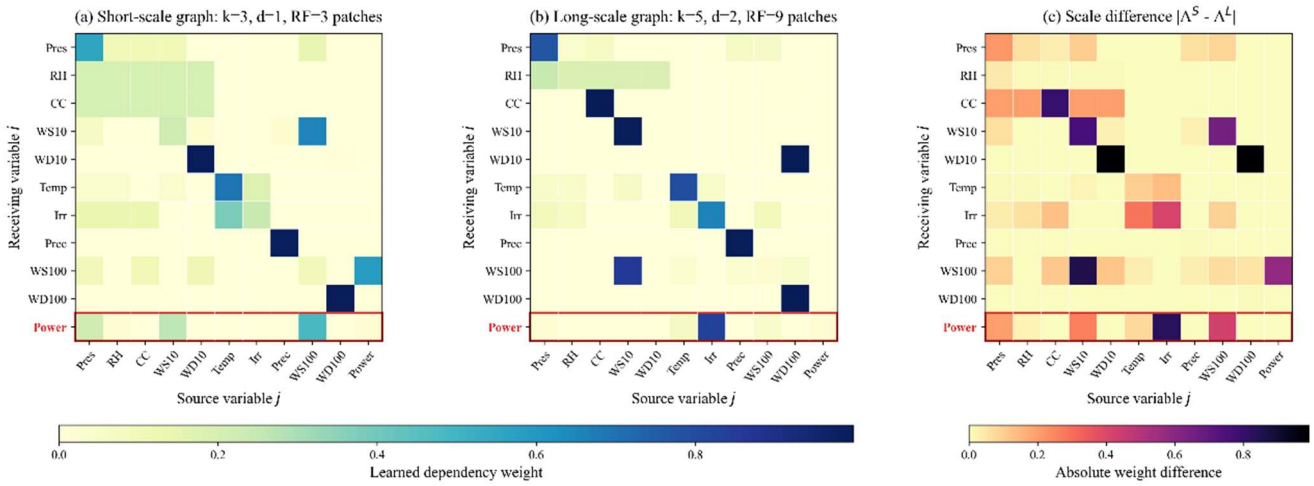


Figure 13. Multi-scale inter-variable dependency matrices learned by MSTGraph: (**a**) short-scale graph; (**b**) long-scale graph; and (**c**) absolute differences between the adjacency weights at the two scales. The horizontal axis represents the source variables, and the vertical axis represents the receiving variables. The red boxes highlight the row in which Power is the receiving variable. The matrix elements represent the learned dependency weights rather than correlation coefficients or causal relationships.

To further evaluate the model's dynamic tracking capability under ramp-intensive conditions, Figure 14 selects a sequence of 200 consecutive test windows from wind farm S2 containing 141 ramp points and compares the forecasts produced by MSTGraph-RAR and its *w/o* RAL variant. Both models capture the overall variation trend of the power series. However, after removing RAL, the forecast trajectory exhibits more pronounced amplitude attenuation during sustained increases, high-power fluctuations, and transitions between local peaks and troughs. Within this interval, incorporating RAL reduces MSE from 349.71 MW^2 to 295.52 MW^2, corresponding to a reduction of 15.50%, while MAE decreases from 15.02 MW to 13.82 MW, representing a reduction of 7.98%. These results indicate that ramp-state auxiliary supervision encourages the shared representations to preserve information associated with rapid regime transitions, thereby improving dynamic tracking performance in ramp-intensive scenarios.

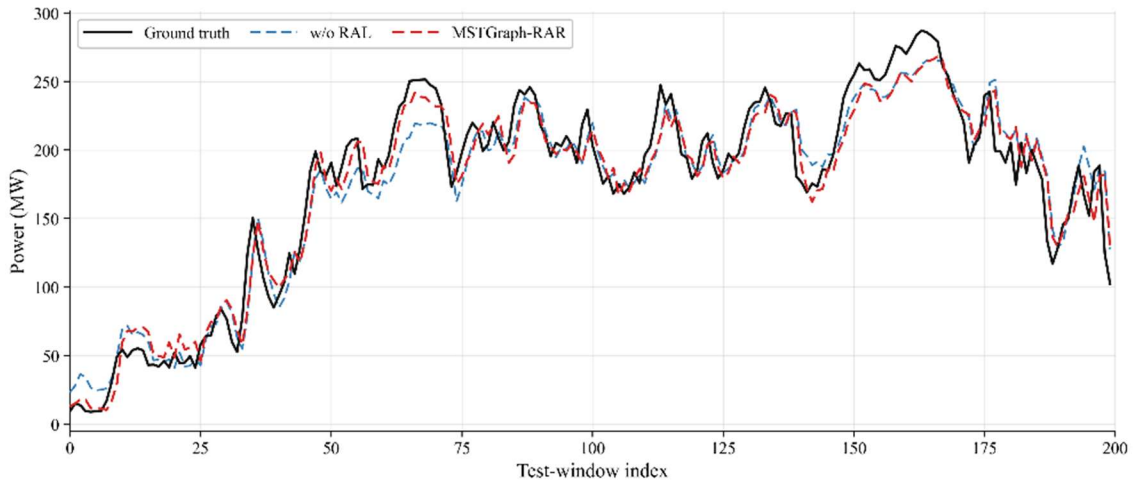


Figure 14. Comparison of power forecasts produced by MSTGraph-RAR and its *w/o* RAL variant over a ramp-intensive sequence of 200 consecutive windows.

To visually assess the practical effect of Ramp-Aware Residual Correction (RRC), Figure 15 compares the forecasts with and without RRC over a continuous 200-point interval. Figure 15a shows that the base forecasts without RRC exhibit clear systematic underestimation during sustained ramps and medium-to-high power regimes. After residual correction, the forecast trajectory more closely follows the observed

power series, with improved tracking of local peaks and high-power plateaus. Figure 15b further presents the pointwise improvement in absolute error, where positive values indicate error reductions achieved by RRC, whereas negative values indicate increased errors after correction. Positive improvements dominate the interval: RRC reduces the forecasting error at 189 of the 200 test points, corresponding to 94.5%.

Overall, incorporating RRC reduces the MSE over this interval from 1085.38 MW^2 to 302.23 MW^2, corresponding to a reduction of 72.15%. The MAE decreases from 26.75 MW to 12.98 MW, representing a reduction of 51.47%. These results indicate that RRC effectively compensates for systematic residual errors produced by the base model during ramp events and medium-to-high power regimes, thereby enhancing the model's ability to track rapid power variations under complex operating conditions.

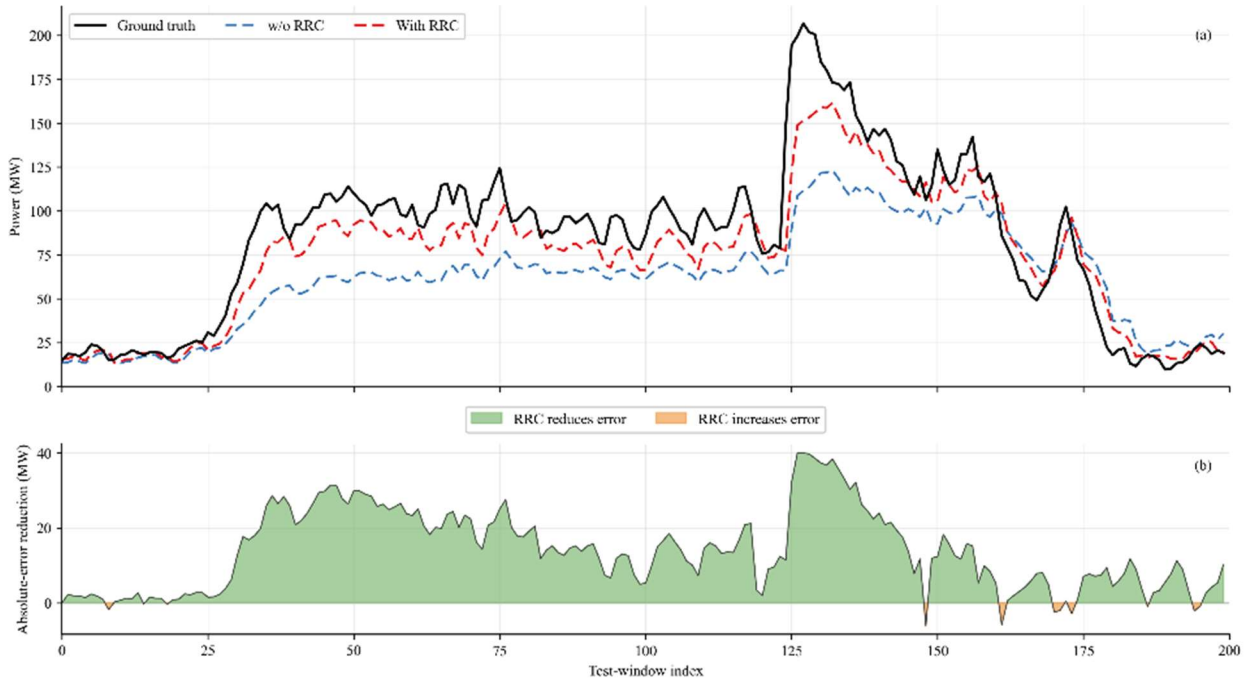


Figure 15. Comparison of power forecasts with and without RRC and the resulting pointwise error improvements: (**a**) power forecasts before and after correction; (**b**) pointwise improvement in absolute error.

4.3. Cross-Farm Generalization Experiments

Beyond forecasting accuracy, cross-farm generalization is important for assessing the practical applicability of forecasting models. Offshore wind farms differ in power-output scale, meteorological conditions, and wind-to-power conversion characteristics. Therefore, performance on farms included in training may not fully reflect the model's transferability to unseen farms. To evaluate this capability, a leave-one-site-out (LOSO) protocol is adopted.

In each LOSO fold, one wind farm is held out, while the remaining four farms retain their chronological 70%/10%/20% data splits. Model parameters, standardization statistics, ramp thresholds, and loss-dependent quantities are determined exclusively from source-farm training data. Source-farm validation data are used for early stopping and checkpoint selection. The trained model is evaluated directly on the held-out farm without target-farm fine-tuning or using its data for parameter estimation, hyperparameter selection, or model selection. Repeating this procedure for all five farms yields five independent transfer tasks.

Evaluation is restricted to the final 20% of observations from each held-out farm, spanning 24 January 2023 at 09:45 to 30 April 2023 at 23:45, consistent with the source-farm test period. Each forecasting sample uses 192 historical steps to predict the subsequent 12, 24, 48, or 96 steps, with both the input and prediction windows entirely within the held-out test subset.

Table 8 reports the MSE of both models separately for each held-out wind farm and forecasting horizon. MSTGraph-RAR achieves lower MSE than PatchTST in 19 of the 20 farm–horizon combinations. Positive relative improvements range from 0.325% to 6.361%.

Table 8. MSE for each held-out farm and forecasting horizon under LOSO.

Held-Out Farm	Forecasting Horizon	PatchTST	MSTGraph-RAR	Relative Improvement
S1	12	0.005622	0.005557	1.156%
	24	0.010465	0.010411	0.516%
	48	0.016381	0.016150	1.410%
	96	0.021261	0.021192	0.325%
S2	12	0.186657	0.180010	3.561%
	24	0.344861	0.324709	5.844%
	48	0.604486	0.566034	6.361%
	96	0.958527	0.932393	2.726%
S3	12	0.007917	0.007588	4.156%
	24	0.014555	0.014359	1.347%
	48	0.024353	0.022895	5.987%
	96	0.032746	0.032098	1.979%
S4	12	0.021598	0.021243	1.644%
	24	0.040846	0.040488	0.876%
	48	0.074590	0.080591	−8.045%
	96	0.127705	0.123990	2.909%
S5	12	0.007678	0.007587	1.185%
	24	0.015100	0.014971	0.854%
	48	0.023640	0.022991	2.745%
	96	0.035071	0.034478	1.691%

Figure 16 presents the MSE improvement of MSTGraph-RAR over PatchTST for each held-out wind farm across different forecasting horizons. The results reveal distinct transfer patterns among the target farms. Most farms exhibit relatively consistent positive improvements across the majority of forecasting horizons, whereas wind farm S4 shows noticeable negative transfer only at the 48-step horizon.

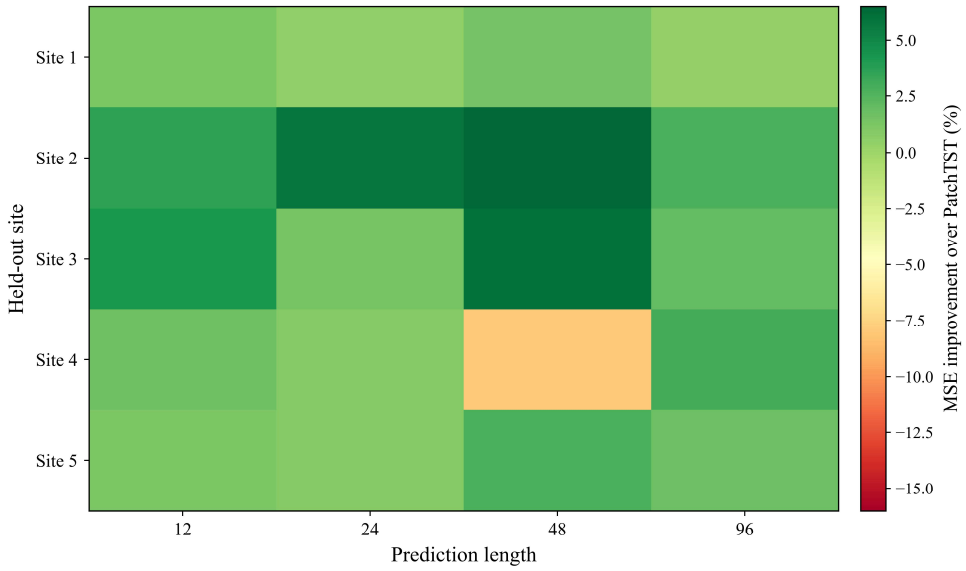


Figure 16. Relative MSE improvements of MSTGraph-RAR over PatchTST for each held-out wind farm and forecasting horizon under the LOSO protocol.

Overall, MSTGraph-RAR improves zero-shot transfer performance across wind farms on average but does not guarantee consistent gains for every unseen offshore wind farm. These findings further suggest that addressing distribution shifts among wind farms remains an important direction for practical multi-farm deployment.

4.4. Ramp-Threshold Sensitivity Analysis

To examine the effect of the ramp-event definition on forecasting accuracy, four threshold configurations are compared: Pooled Q85, Pooled Q90, Pooled Q95, and Farm-specific Q90. The pooled thresholds are estimated from the absolute power increments in the combined training subsets of S1–S5 and correspond to power changes of 3.84, 5.44, and 9.42 MW over a 15-min interval, respectively. Farm-specific Q90 determines the threshold from each farm's own training subset while retaining the common standardization procedure across farms. Pooled Q90 serves as the reference configuration, with the data splits, model architecture, and optimization hyperparameters held constant across all configurations. Table 9 reports the mean and sample standard deviation of test-set MSE across three random seeds.

Table 9. Threshold sensitivity of MSTGraph-RAR across forecasting horizons.

Metric	Threshold Setting	12	24	48	96
MSE	Pooled Q85	0.0768 ± 0.0017	0.1458 ± 0.0023	0.2609 ± 0.0106	0.4078 ± 0.0119
MSE	Pooled Q90	0.0764 ± 0.0024	0.1424 ± 0.0016	0.2498 ± 0.0040	0.3998 ± 0.0027
MSE	Pooled Q95	0.0801 ± 0.0018	0.1420 ± 0.0024	0.2543 ± 0.0006	0.3982 ± 0.0150
MSE	Farm-specific Q90	0.0790 ± 0.0047	0.1458 ± 0.0007	0.2603 ± 0.0107	0.4191 ± 0.0127

Table 9 shows that aggregate forecasting accuracy remains relatively stable as the pooled threshold varies from Q85 to Q95, corresponding to absolute 15-min power changes of 3.84–9.42 MW. The maximum increases in mean MSE relative to Q90 are 4.84%, 2.39%, 4.44%, and 2.00% at the 12-, 24-, 48-, and 96-step horizons, respectively. These bounded changes indicate limited sensitivity to the pooled ramp threshold within the examined range.

Lower thresholds assign ramp labels to a broader set of power changes, whereas higher thresholds restrict ramp supervision to rarer, larger changes. The horizon-dependent responses may reflect a balance between supervision coverage and selectivity. Q90 achieves the lowest mean MSE at 12 and 48 steps, while its mean MSE exceeds the lowest observed values by only 0.28% and 0.40% at 24 and 96 steps, respectively. At 96 steps, Q90 also exhibits lower variation across seeds than Q95, with standard deviations of 0.0027 and 0.0150, respectively. Together, these findings support Q90 as a reasonable default that maintains competitive accuracy across forecasting horizons.

The threshold-construction strategy also affects forecasting performance. Compared with Farm-specific Q90, Pooled Q90 reduces aggregate mean MSE by 3.29%, 2.33%, 4.03%, and 4.61% across the four horizons, respectively. Together with the limited variations observed among Pooled Q85, Q90, and Q95, these results indicate that MSTGraph-RAR is relatively insensitive to threshold selection within the evaluated settings.

5. Conclusions

This study proposes the MSTGraph-RAR framework to address the coupled modeling of multi-scale temporal dynamics and inter-variable dependencies, the insufficient representation of rapid local power variations, and the increasing residual errors over extended forecasting horizons in multivariate multi-step offshore wind power forecasting. The framework is systematically evaluated using data from five real-world offshore wind farms across four forecasting horizons, together with ablation studies, ramp-intensive scenario analyses, and cross-farm generalization experiments. Overall, MSTGraph-RAR achieves the best

performance in the 12-, 24-, 48-, and 96-step forecasting tasks. Compared with the competing models, it delivers maximum reductions of 36.13%, 24.80%, and 19.71% in MSE, MAE, and RMSE, respectively, together with an improvement of up to 0.13 in R^2. These results demonstrate that the proposed framework maintains consistent forecasting advantages across different horizons.

Further analysis shows that MSTGraph effectively complements channel-independent temporal representations by jointly learning temporal patterns and inter-variable dependencies at different scales. The short- and long-scale branches learn distinct interaction structures, indicating that the dependencies between offshore wind power and meteorological variables exhibit scale-dependent characteristics. Consequently, a single temporal scale or fixed inter-variable relationships may be insufficient to fully characterize complex multivariate dynamics. RAL introduces future power-change states as auxiliary supervision, enhancing the ability of the shared representations to distinguish rapid increases, decreases, and transitions between local peaks and troughs. In the ramp-intensive scenario, RAL reduces MSE by 15.50%, indicating that ramp-related state information can effectively contribute to representation learning for continuous multi-step power forecasting.

Building on this, RRC adaptively assigns the correction intensity to individual future forecasting steps and uses bounded residuals to prevent excessive correction. The experimental results show that this residual correction further mitigates forecasting bias while reducing unnecessary perturbations to already reliable base forecasts. In the cross-farm LOSO experiments, MSTGraph-RAR achieves lower MSE than PatchTST in 19 of the 20 farm–horizon combinations, demonstrating its transfer potential across unseen wind farms.

Overall, MSTGraph-RAR provides an effective solution with demonstrated generalization potential for multivariate multi-step offshore wind power forecasting under complex non-stationary conditions. The present analysis has two main limitations. First, the present study focuses primarily on deterministic power forecasting and does not explicitly model forecasting uncertainty. Second, the inter-variable dependencies learned by MSTGraph are predominantly data-driven, and their physical interpretations require further validation against marine meteorological mechanisms. Future research will therefore investigate probabilistic forecasting and the integration of physical knowledge, drawing on numerical studies of marine environmental effects on offshore structures [35], to further improve model reliability and interpretability.

Statement of the Use of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT (GPT-5.5; OpenAI) in order to assist with grammatical correction. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Author Contributions

Conceptualization, W.X. and F.S.; Methodology, L.Z.; Software, L.Z.; Validation, M.G. and S.Y.; Formal Analysis, Y.W.; Investigation, Y.Z.; Data Curation, Y.T.; Writing—Original Draft Preparation, L.Z.; Writing—Review & Editing, M.G. and W.X.

Ethics Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The processed data supporting the findings are available from the corresponding author upon reasonable request.

Funding

This study was supported by the project "Development of Wind Power Prediction Model Based on Deep Learning and Its Application in Wind Farm Site Planning" (H20240396).

Declaration of Competing Interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Lyu Y, Chen H, Yazan DM. A comprehensive framework for data challenges and intelligent resource optimization for offshore wind energy. *Sustain. Energy Technol. Assess.* **2026**, *87*, 104888. DOI:10.1016/j.seta.2026.104888
2. Wang Q, Xu F, He J, Luo K, Fan J. A new fusion model for enhanced ultra-short-term offshore wind power forecasting. *Renew. Energy* **2026**, *256*, 123876. DOI:10.1016/j.renene.2025.123876
3. Liu Y, Zhu W, Zhao B, Zhang Y, Guo L, Lv H, et al. A pattern-aware two-stage power forecasting method for offshore wind farms based on dynamic feature decoupling. *Appl. Energy* **2026**, *414*, 127879. DOI:10.1016/j.apenergy.2026.127879
4. Liu D, Yan Y, Chen S, Wang B, Li X, Zhang H. Wind power forecasting with deep representation: A survey. *Appl. Energy* **2026**, *421*, 128106. DOI:10.1016/j.apenergy.2026.128106
5. Xu F, Wang Q, Du X, Luo K, Fan J. A novel hybrid short-term and ultra-short-term wind power forecasting method based on Weather Research and Forecasting: WRF-iTransformer-PSO. *Energy* **2026**, *353*, 140955. DOI:10.1016/j.energy.2026.140955
6. Cai C, Shi Q, Jin Y, Hua M, Yang T, Hou S. Offshore wind power forecasting with wind-regime clustering and multi-scale feature learning. *Int. J. Electr. Power Energy Syst.* **2026**, *174*, 111553. DOI:10.1016/j.ijepes.2025.111553
7. Zhao W, Zhao J, Shen X, Liu J, Fan S. A two-stage uncertainty forecasting method for offshore wind power based on FEDformer and SDCDM. *Energy* **2026**, *348*, 140452. DOI:10.1016/j.energy.2026.140452
8. Song W, Yan J, Han S, Liu S, Wang H, Dai Q, et al. A multi-task spatio-temporal fusion network for offshore wind power ramp events forecasting. *Renew. Energy* **2024**, *237*, 121774. DOI:10.1016/j.renene.2024.121774
9. Morales-Hernández A, De Caro F, Paldino GM, Tribel P, Vaccaro A, Bontempi G. A direct classification approach for reliable wind ramp event forecasting under severe class imbalance. *Electr. Power Syst. Res.* **2026**, *257*, 112951. DOI:10.1016/j.epsr.2026.112951
10. Ren G, Wan J, Wang Y, Yao K, Fu J, Yu J. A direct prediction method for wind power ramp events considering the class imbalanced problem. *Energy Sci. Eng.* **2023**, *11*, 1705–1715. DOI:10.1002/ese3.1415
11. Feng X, Guo D, Ye M. DMoE-LLM: A dual-branch mixture-of-experts framework with large language models for wind power forecasting. *Expert Syst. Appl.* **2026**, *325*, 132637. DOI:10.1016/j.eswa.2026.132637
12. Zhang P, Li C, Qian C, Huang H, Wang Y, Tang R. A wind power forecasting method based on GWS-STNet spatio-temporal meteorological prediction. *Energy* **2026**, *350*, 140714. DOI:10.1016/j.energy.2026.140714
13. Sun K, Yang D, Wang D, Zhang F. Development of a physics-guided bidirectional long short-term memory for wind power forecasting. *Eng. Appl. Artif. Intell.* **2026**, *167*, 113798. DOI:10.1016/j.engappai.2026.113798
14. Zhou T, Ma Z, Wen Q, Wang X, Sun L, Jin R. FEDformer: Frequency enhanced decomposed Transformer for long-term series forecasting. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 27268–27286.
15. Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, et al. iTransformer: Inverted Transformers are effective for time series forecasting. In Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024. Available online: https://openreview.net/forum?id=JePfAI8fah (accessed on 1 September 2026).
16. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: Long-term forecasting with Transformers. In Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=Jbdc0vTOcol (accessed on 1 September 2026).

17. Kim D, Ryu G, Moon C, Kim B. Accuracy of a short-term wind power forecasting model based on deep learning using LiDAR-SCADA integration: A case study of the 400-MW Anholt offshore wind farm. *Appl. Energy* **2024**, *373*, 123882. DOI:10.1016/j.apenergy.2024.123882
18. Sun Y, Zhou Q, Sun L, Sun L, Kang J, Li H. CNN–LSTM–AM: A power prediction model for offshore wind turbines. *Ocean Eng.* **2024**, *301*, 117598. DOI:10.1016/j.oceaneng.2024.117598
19. Huang J, Qin R. Elman neural network considering dynamic time delay estimation for short-term forecasting of offshore wind power. *Appl. Energy* **2024**, *358*, 122671. DOI:10.1016/j.apenergy.2024.122671
20. Fang L, He B, Yu S. A modular multi-step forecasting method for offshore wind power clusters. *Appl. Energy* **2025**, *380*, 125060. DOI:10.1016/j.apenergy.2024.125060
21. Xiong Y, Wang W. Offshore wind power multi-step forecasting based on multi-scale attention mechanism fusion network. *Eng. Appl. Artif. Intell.* **2025**, *161*, 112026. DOI:10.1016/j.engappai.2025.112026
22. Wang L, Gao J, Li Y, Wang D. A method for ultra-short-term wind power forecasting of large-scale wind farms based on adaptive spatiotemporal graph convolution. *Renew. Energy* **2025**, *249*, 123321. DOI:10.1016/j.renene.2025.123321
23. Wang F, Chen P, Zhen Z, Yin R, Cao C, Zhang Y, et al. Dynamic spatio-temporal correlation and hierarchical directed graph structure based ultra-short-term wind farm cluster power forecasting method. *Appl. Energy* **2022**, *323*, 119579. DOI:10.1016/j.apenergy.2022.119579
24. Xiao L, Meng A, Zhang Q, Song S, Yin H, Luo J. Offshore wind farm cluster power forecasting: An ASTGCN framework with dual-layer co-evolution of loss function and graph construction method. *Energy* **2026**, *347*, 140394. DOI:10.1016/j.energy.2026.140394
25. Chen L, Chen D, Shang Z, Wu B, Zheng C, Wen B, et al. Multi-scale adaptive graph neural network for multivariate time series forecasting. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 10748–10761. DOI:10.1109/TKDE.2023.3268199
26. Cai W, Liang Y, Liu X, Feng J, Wu Y. MSGNet: Learning multi-scale inter-series correlations for multivariate time series forecasting. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 11141–11149. DOI:10.1609/aaai.v38i10.28991
27. Gao Y, Hu S, Chen Y, Khan MF, Cheng X, Yang J. A novel probabilistic wind power forecasting framework integrating similar curve matching mechanism and an enhanced conditional diffusion model. *Appl. Energy* **2026**, *402*, 127031. DOI:10.1016/j.apenergy.2025.127031
28. Li M, Zhang K, Kou M, Ma Y. An offshore wind speed forecasting system based on feature enhancement, deep time series clustering, and extended LSTM. *Energy* **2025**, *333*, 137335. DOI:10.1016/j.energy.2025.137335
29. Ouyang T, Zha X, Qin L, He Y, Tang Z. Prediction of wind power ramp events based on residual correction. *Renew. Energy* **2019**, *136*, 781–792. DOI:10.1016/j.renene.2019.01.049
30. Wu Z, Pan S, Long G, Jiang J, Zhang C. Graph WaveNet for deep spatial-temporal graph modeling. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 1907–1913. DOI:10.24963/ijcai.2019/264
31. Wu Z, Pan S, Long G, Jiang J, Chang X, Zhang C. Connecting the dots: Multivariate time series forecasting with graph neural networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, USA, 23–27 August 2020; pp. 753–763. DOI:10.1145/3394486.3403118
32. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. DOI:10.1609/aaai.v35i12.17325
33. Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=ju_Uqw384Oq (accessed on 1 September 2026).
34. Wang S, Wu H, Shi X, Hu T, Luo H, Ma L, et al. TimeMixer: Decomposable Multi-scale Mixing for Time Series Forecasting. In Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024. Available online: https://openreview.net/forum?id=7oLshfEIC2 (accessed on 1 September 2026).
35. Mohapatra SC, Amouzadrad P, Bispo IBDS, Guedes Soares C. Hydrodynamic Response to Current and Wind on a Large Floating Interconnected Structure. *J. Mar. Sci. Eng.* **2025**, *13*, 63. DOI:10.3390/jmse13010063