

Article

RegimeMoE-PatchTST: Metocean-Conditioned Mixture of PatchTST Experts for Multi-Step Offshore Wind Power Forecasting

Mingju Gong¹, Yaming Wang¹, Minghao Sun¹, Yu Tian¹, Linlin Zhu¹, Siyuan Yin¹, Yilin Zhang¹, Wei Xu^{2,3,*} and Fei Su¹

¹ School of Integrated Circuit Science and Engineering, Tianjin University of Technology, Tianjin 300384, China; gmj790@email.tjut.edu.cn (M.G.); wangyaming@stud.tjut.edu.cn (Y.W.); wyn001221@stud.tjut.edu.cn (M.S.); tianyu@stud.tjut.edu.cn (Y.T.); 13903219569@163.com (L.Z.); yinsiyuan@stud.tjut.edu.cn (S.Y.); zyl2487191829@stud.tjut.edu.cn (Y.Z.); sufei@email.tjut.edu.cn (F.S.)

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: xuwei_tjut@163.com (W.X.)

Received: 29 August 2026; Revised: 28 September 2026; Accepted: 30 September 2026; Available online: 20 October 2026

ABSTRACT: Offshore wind power forecasting is challenging because rapidly varying metocean conditions induce strong non-stationarity and heterogeneous power-generation dynamics. This study proposes RegimeMoE-PatchTST, a metocean-conditioned mixture-of-experts framework for multi-step offshore wind power forecasting. For each historical window, four statistical descriptors of seven metocean variables are extracted to form a 28-dimensional condition vector. Three independently parameterized PatchTST networks serve as forecasting experts, while a gating network maps the condition vector to sample-dependent weights and softly fuses their multi-step predictions. The entire framework is trained end-to-end without predefined regime labels. Experiments on two offshore wind turbine datasets at 2-, 4-, 8-, and 16-h horizons show that the proposed model achieves the lowest mean squared error (MSE) in all eight turbine–horizon settings and reduces average MSE by 3.69% relative to PatchTST. Ablation results verify the importance of both metocean-conditioned routing and dynamic expert fusion, while regime-wise evaluation shows MSE reductions of 7.7% and 8.1% under high-wind and high-turbulence conditions. These results demonstrate improved adaptability and forecasting accuracy under heterogeneous offshore operating conditions.

Keywords: Offshore wind power forecasting; PatchTST; Mixture of experts; Metocean conditions; Multi-step forecasting

1. Introduction

Wind energy has become one of the most important renewable energy sources in the transition toward low-carbon power systems. Compared with onshore wind energy, offshore wind energy offers several



advantages, including abundant wind resources, relatively stable wind conditions, large-scale development potential, and limited occupation of onshore land. However, the stochastic, intermittent, and nonlinear nature of wind generation introduces considerable uncertainty into power-system operation [1–5]. Offshore wind power forecasting is particularly challenging because turbine power output is jointly affected by rapidly changing wind conditions, atmospheric properties, turbine operating states, and complex environmental interactions [6–8]. In related offshore engineering research, Mohapatra et al. [9] examined the hydrodynamic response of a large interconnected floating structure under waves, currents, and wind. Accurate offshore wind power forecasting is therefore essential for power-system scheduling, reserve allocation, wind-power integration, energy trading, and wind-farm operation and maintenance.

Data-driven approaches have become increasingly important for wind power forecasting because of their ability to learn nonlinear relationships directly from historical observations [1–3]. Early deep-learning-based forecasting studies mainly employed convolutional neural networks (CNNs), recurrent neural networks, long short-term memory (LSTM) networks, and gated recurrent units (GRUs) to extract local patterns and sequential dependencies. More recently, hybrid architectures incorporating signal decomposition, attention mechanisms, numerical weather prediction (NWP), transfer learning, and probabilistic modeling have been developed to improve forecasting accuracy and robustness [10–18]. For example, Ma and Mei [10] combined multiple deep neural components with an attention mechanism to enhance temporal feature extraction. Liu et al. [11] investigated NWP enhancement through rank ensemble and probabilistic fluctuation awareness, while a subsequent study introduced series-wise attention to address the time-varying temporal lag between NWP wind speed and actual measurements [12]. Hu et al. [13] proposed temporal collaborative attention to model interactions among temporal and multivariate features, and Dai et al. [14] developed an online-learning-enabled self-attention architecture to adapt to time-varying wind-power patterns. Transfer learning and probabilistic approaches have been further explored to address distribution shifts, limited historical observations, and forecasting uncertainty [15–18]. In offshore applications, Lin and Liu [6] demonstrated the value of high-frequency supervisory control and data acquisition (SCADA) measurements for turbine-level forecasting, Hanifi et al. [7] combined wavelet packet decomposition with optimized CNN and LSTM models, and Kim et al. [8] showed that integrating light detection and ranging (LiDAR) and SCADA information can improve short-term offshore wind power prediction. These studies demonstrate the importance of both temporal representation learning and environmental information for accurate wind power forecasting.

Despite these advances, recurrent and multi-stage hybrid architectures still present several limitations. Recurrent networks process observations sequentially, restricting computational parallelism and potentially making long-range dependency learning difficult. Hybrid models involving signal decomposition, multiple preprocessing stages, or complex feature-fusion procedures can achieve high forecasting accuracy, but they often increase the computational and implementation complexity of the overall forecasting pipeline [3,7,10]. These limitations have motivated the increasing use of Transformer-based architectures for time-series forecasting. Temporal Fusion Transformer introduced attention-based multi-horizon forecasting with variable selection and gating mechanisms [19]. Informer reduced the computational burden of long-sequence attention through ProbSparse self-attention [20], whereas Autoformer incorporated progressive decomposition and an auto-correlation mechanism for long-term forecasting [21]. FEDformer further introduced frequency-enhanced representations to capture global temporal patterns [22]. At the same time, the strong performance of simple linear models reported by Zeng et al. [23] motivated reconsideration of the inductive biases required for time-series forecasting. Subsequent architectures such as TimesNet [24], TiDE [25], and iTransformer [26] explored two-dimensional temporal variation modeling, dense encoder–decoder structures, and variable-oriented attention, respectively.

Among these architectures, PatchTST has attracted particular attention because of its patch-based and channel-independent design [27]. Instead of treating individual time points as separate tokens, PatchTST

divides each variable sequence into local temporal patches and feeds the patch representations into a Transformer encoder. This design preserves local temporal semantics, reduces the effective attention sequence length, and enables the model to utilize a longer historical context at manageable computational cost. Channel-independent encoding further prevents premature mixing of heterogeneous variables while allowing shared temporal representations to be learned across channels [27]. These characteristics make PatchTST well suited to direct multi-step forecasting. Transformer-based models have also increasingly been adopted for wind power prediction. Wang et al. [28], for example, incorporated static information into a Transformer-based framework for very-short-term forecasting, while Xu et al. [29] combined PatchTST with a GRU decoder and refined NWP information for multi-step wind power forecasting. In addition, spatial graph modeling [30] and physics-constrained deep learning [31] have demonstrated that incorporating turbine interactions and domain knowledge of wind power can further enhance forecasting reliability.

However, most existing forecasting architectures primarily focus on improving temporal representation, cross-variable interaction, spatial dependency modeling, or multi-source information fusion. Comparatively less attention has been paid to whether the forecasting function itself should adapt to heterogeneous environmental operating conditions. This issue is particularly relevant for offshore wind turbines. Power-generation dynamics under low-, medium-, and high-wind conditions can differ substantially because of the nonlinear turbine power curve and turbine control behavior. Short-term fluctuations can be stronger during periods of high turbulence than under relatively stable atmospheric conditions, while variations in wind direction, air density, humidity, and other environmental variables can further modify the relationship between historical observations and future power output. Previous studies on wind power curves have also shown that no single modeling strategy consistently performs best under all environmental conditions [4]. Therefore, applying one shared set of forecasting parameters to all historical windows may restrict the model's ability to represent heterogeneous offshore operating states.

Recent studies have begun to explicitly consider such operating-regime heterogeneity. Cai et al. [32] proposed an offshore wind power forecasting framework that combines wind-direction-based pre-classification and K-means clustering to organize historical observations into distinct meteorological regimes before multi-scale feature learning. Isik and Yalcinkaya [33] similarly developed a regime-aware forecasting pipeline based on multi-altitude meteorological features, unsupervised regime identification, and regime-conditioned forecasting models. These studies demonstrate that separating or conditioning forecasting behavior according to atmospheric states can improve prediction performance. Nevertheless, explicit regime-based approaches generally require a preprocessing or clustering stage to construct discrete regime labels. Their forecasting behavior consequently depends, at least partly, on the definition and separation of predefined or externally identified regimes. For highly dynamic offshore conditions, transitions between operating states may be gradual rather than strictly discrete. A desirable alternative is therefore to allow the forecasting strategy to change continuously according to the environmental state of each input window, without requiring manually specified or pre-clustered regime labels.

Mixture-of-experts (MoE) architectures provide a natural framework for this purpose. The original mixture-of-local-experts formulation divides heterogeneous samples among multiple specialized predictors through a trainable gating mechanism [34]. Sparse MoE architectures subsequently demonstrated that conditional expert activation can substantially increase model capacity while controlling computational cost [35]. More recently, this concept has been extended to time-series forecasting. Time-MoE employs sparsely activated experts to construct billion-scale time-series foundation models [36], while Moirai-MoE introduces automatic token-level specialization to capture heterogeneous temporal patterns without relying on manually defined frequency groups [37]. These developments demonstrate the ability of MoE architectures to model heterogeneous time-series distributions through expert specialization and adaptive routing.

MoE has also recently been introduced into wind power forecasting. Hossain et al. [38] proposed a mathematically guided adaptive MoE framework for ultra-short-term wind power prediction, in which

multiple heterogeneous forecasting models are combined using a context-aware gating mechanism assisted by statistical information such as volatility and trend. Zhao et al. [39] developed a physics-aware dynamic-graph MoE framework that combines temporal encoding, dynamic turbine-interaction modeling, and gated fusion between physics-informed and data-driven predictors. These studies confirm the potential of dynamic expert selection for non-stationary wind power forecasting. However, their objectives and architectural designs differ from the problem considered in this study. Existing wind-power MoE methods mainly emphasize heterogeneous model ensembles, recent-performance-guided gating, physics–data fusion, or wind-farm spatial interactions. The use of a compact window-level environmental condition representation to continuously coordinate multiple independently parameterized patch-based temporal experts for multi-step offshore turbine forecasting remains insufficiently investigated.

To address this gap, this study proposes RegimeMoE-PatchTST, a metocean-conditioned mixture-of-experts framework for multi-step offshore wind power forecasting. Instead of explicitly assigning input samples to predefined operating regimes, the proposed framework summarizes the environmental state of each historical window through a compact condition representation. Specifically, four statistical descriptors—the mean, standard deviation, latest observation, and temporal change—are extracted from seven environmental variables to form a 28-dimensional condition vector. In parallel, three PatchTST networks with identical architectures but independently learned parameters are employed as forecasting experts to model the complete multivariate historical sequence. A regime-aware gating network maps the condition vector to sample-dependent expert weights and softly combines the complete multi-step forecasts produced by the experts. The gating network and PatchTST experts are jointly optimized using the final forecasting objective, allowing expert specialization and routing behavior to emerge directly from the data without requiring predefined regime labels.

The main contributions of this study are summarized as follows:

1. A compact metocean condition representation is developed for offshore wind power forecasting. Four statistical descriptors are extracted from seven environmental variables to construct a 28-dimensional condition vector. The representation characterizes the average operating level, local variability, latest state, and overall temporal tendency of each historical input window, thereby providing physically meaningful condition information for expert routing.
2. A metocean-conditioned mixture of PatchTST experts is proposed for multi-step offshore wind power forecasting. Three PatchTST predictors with identical architectures but independently learned parameters are jointly trained as forecasting experts. The independently parameterized branches enable the model to learn differentiated temporal representations, while the gating network dynamically adjusts their contributions based on the environmental state of each sample.
3. A sample-dependent soft-routing mechanism without predefined regime labels is developed and systematically evaluated. Rather than assigning samples to manually defined wind regimes, the model learns continuous expert combinations directly from the forecasting objective. Experiments on two offshore wind turbine datasets with forecasting horizons of 2, 4, 8, and 16 h show that RegimeMoE-PatchTST achieves the lowest mean squared error (MSE) in all evaluated turbine–horizon combinations and reduces the average MSE by 3.69% relative to the PatchTST backbone. Ablation experiments demonstrate that consistent improvements require the joint use of environmental condition information and dynamic expert routing. Furthermore, regime-wise analysis shows MSE reductions of 7.7% and 8.1% under high-wind and high-turbulence conditions, respectively, indicating that the proposed mechanism provides greater benefits under challenging offshore operating states.

The remainder of this paper is organized as follows. Section 2 presents the proposed RegimeMoE-PatchTST framework, including the PatchTST forecasting experts, metocean condition representation, regime-aware routing mechanism, and training objective. Section 3 describes the offshore wind turbine datasets, preprocessing procedures, baseline models, evaluation metrics, and experimental settings. Section

4 reports the main forecasting results, ablation studies, expert-usage analysis, and metocean-regime-wise evaluation. Finally, Section 5 concludes the paper and discusses potential directions for future research.

2. Proposed Method

This section presents the proposed RegimeMoE-PatchTST framework for multi-step offshore wind power forecasting. The framework combines multiple independently parameterized PatchTST forecasting experts with a metocean-conditioned routing mechanism. The PatchTST experts learn temporal dependencies from the complete historical sequence, while a compact metocean condition vector is constructed from each input window to characterize its operating condition. A gating network then generates sample-dependent expert weights, and the final forecast is obtained by dynamically combining the expert predictions. In this manner, the forecasting strategy can adapt to heterogeneous offshore operating conditions without requiring predefined regime labels.

2.1. Problem Formulation

Equations (1)–(4) adapt the historical-window forecasting formulation of PatchTST [27] to the wind-power prediction task. At time step t , let the multivariate observation vector be

$$x_t = [m_t, p_t] \in R^C \quad (1)$$

where $m_t \in R^7$ denotes the seven metocean variables, including wind speed, adjusted wind speed, the sine and cosine components of wind direction, air density, humidity, and turbulence intensity. The scalar p_t denotes the normalized wind power. Therefore, each time step contains $C = 8$ variables.

Given a historical input window of length L , the input sequence is expressed as

$$X_t = [x_{t-L+1}, x_{t-L+2}, \dots, x_t]^T \in R^{L \times C} \quad (2)$$

For a forecasting horizon of H time steps, the prediction target is defined as

$$y_t = [p_{t+1}, p_{t+2}, \dots, p_{t+H}]^T \in R^H \quad (3)$$

Accordingly, the forecasting task can be formulated as learning the mapping

$$\hat{y}_t = F_\theta(X_t) \quad (4)$$

where $F_\theta(\cdot)$ denotes the forecasting model parameterized by θ , and $\hat{y}_t \in R^H$ denotes the predicted future power sequence.

A direct multi-step forecasting strategy is adopted, in which the complete H – step future power sequence is generated in a single forward pass rather than recursively. This formulation avoids the accumulation of preceding prediction errors and enables the model to jointly capture dependencies among different future time steps.

2.2. Overall Architecture

As illustrated in Figure 1, the proposed RegimeMoE-PatchTST framework consists of three main components: multiple independently parameterized PatchTST forecasting experts, a metocean condition representation module, and a metocean-conditioned gating and fusion module. Given the same historical sequence, the PatchTST experts independently generate multi-step forecasts, while the metocean variables are summarized into a compact condition vector for expert routing.

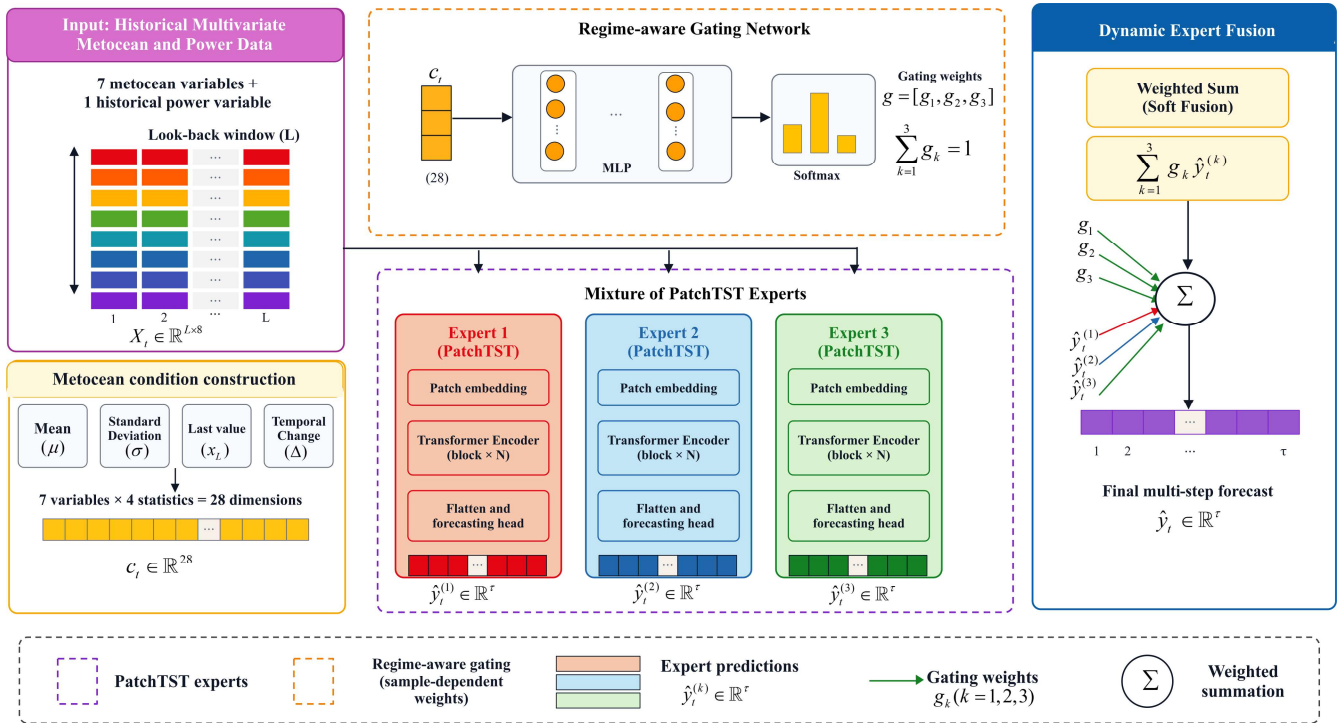


Figure 1. Overall architecture of RegimeMoE-PatchTST.

Let K denote the number of forecasting experts, with $K = 3$ in this study. Given the historical sequence X_t , the prediction produced by the k -th PatchTST expert [27] is expressed as

$$\hat{y}_t^{(k)} = f_{\theta_k}(X_t), \quad k = 1, \dots, K \quad (5)$$

where $f_{\theta_k}(\cdot)$ denotes the k -th expert parameterized by θ_k . Although all experts share the same network architecture and hyperparameter configuration, their parameters are learned independently, allowing different experts to develop differentiated temporal forecasting patterns.

The metocean subsequence and condition mapping in Equations (6) and (7) are defined for this study. In parallel, the seven metocean variables are extracted from X_t to form the metocean subsequence

$$M_t \in \mathbb{R}^{L \times 7} \quad (6)$$

A condition representation function $\Phi(\cdot)$ maps the metocean subsequence to a compact condition vector

$$c_t = \Phi(M_t) \in \mathbb{R}^{28} \quad (7)$$

The softmax gating in Equations (8) and (9) and the weighted aggregation in Equation (10) follow the mixture-of-experts formulation [34]. Based on this representation, the gating network generates the expert-weight vector

$$g_t = G_\psi(c_t) = [g_{t,1}, g_{t,2}, \dots, g_{t,K}]^T \quad (8)$$

where the weights satisfy

$$g_{t,k} \geq 0, \quad \sum_{k=1}^K g_{t,k} = 1 \quad (9)$$

The final prediction is obtained through weighted fusion:

$$\hat{y}_t = \sum_{k=1}^K g_{t,k} \hat{y}_t^{(k)} \quad (10)$$

Because the routing weights are computed from the metocean condition of the current historical window, different samples can employ different expert combinations. In this study, the term regime does not refer to a manually predefined or externally identified discrete meteorological category. Instead, it denotes a latent operating state that varies continuously with the metocean conditions in each historical window and is reflected in the learned expert-weight distribution.

2.3. PatchTST Expert Forecasting

The internal architecture of each PatchTST forecasting expert is illustrated in Figure 2. Each expert adopts PatchTST as its temporal forecasting backbone. Instead of treating individual time points as separate tokens, PatchTST divides each variable sequence into local temporal patches and performs channel-independent Transformer encoding. This design preserves local temporal patterns while reducing the effective sequence length, thereby facilitating the modeling of long-range temporal dependencies.

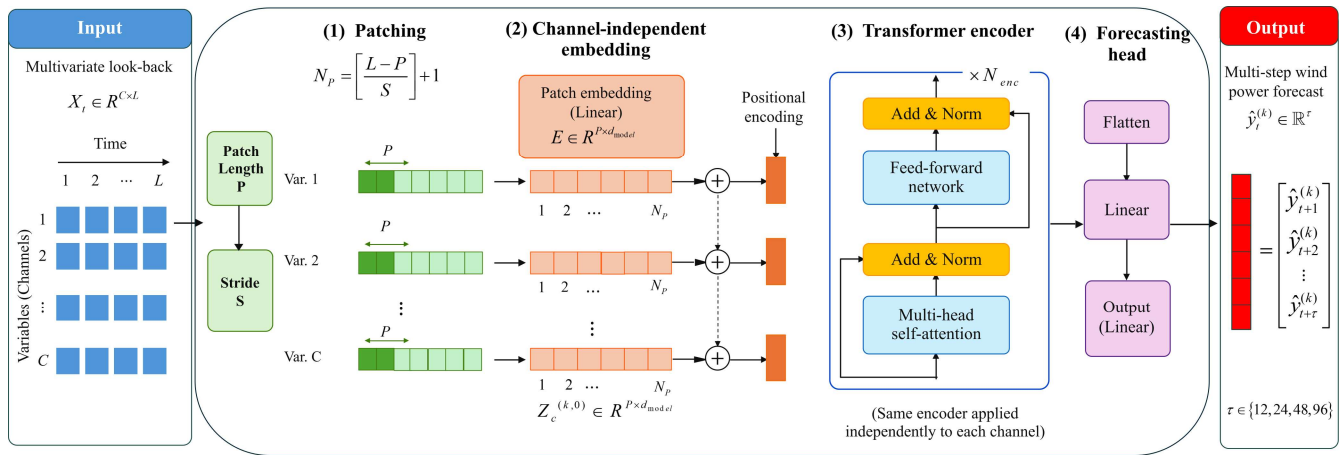


Figure 2. Internal architecture of a PatchTST forecasting expert.

For the c -th input variable, its historical observations within the current input window are denoted by

$$x_t^{(c)} = [x_1^{(c)}, x_2^{(c)}, \dots, x_L^{(c)}] \in R^L, \quad c = 1, \dots, C \quad (11)$$

Given a patch length P and a stride S , the sequence is divided into local temporal patches. The n -th patch is represented as

$$p_{c,n} = [x_{(n-1)S+1}^{(c)}, \dots, x_{(n-1)S+P}^{(c)}] \in R^P \quad (12)$$

The corresponding number of patches is

$$N_p = \left\lfloor \frac{L-P}{S} \right\rfloor + 1 \quad (13)$$

When $S < P$, adjacent patches overlap, which helps preserve temporal information around patch boundaries.

Each patch is projected into a latent representation and combined with positional information before being processed by the Transformer encoder. The encoded representations are subsequently mapped by the forecasting head of the k -th expert to the future power sequence:

$$\hat{y}_t^{(k)} = H_{\theta_k} \left(E_{\theta_k} (X_t) \right) \in R^H \quad (14)$$

where $E_{\theta_k}(\cdot)$ denotes the temporal encoder of the k -th expert and $H_{\theta_k}(\cdot)$ denotes its forecasting head.

PatchTST adopts channel-independent temporal modeling, in which different variables are patched and encoded separately. This design preserves variable-specific temporal characteristics and avoids premature mixing among heterogeneous input variables. The three experts share the same network topology and hyperparameter settings but maintain independent patch projection layers, Transformer encoders, and forecasting heads. Consequently, they have the capacity to learn differentiated temporal forecasting patterns. Their outputs are not simply averaged; instead, their relative contributions are dynamically determined by the metocean-conditioned routing mechanism described in Section 2.5.

2.4. Metocean Condition Representation

As shown in Figure 3, the metocean condition representation is constructed by extracting four statistical descriptors from each of the seven metocean variables within the historical input window. The purpose of this module is to provide the gating network with a compact window-level description of the current operating condition rather than another high-dimensional temporal representation.

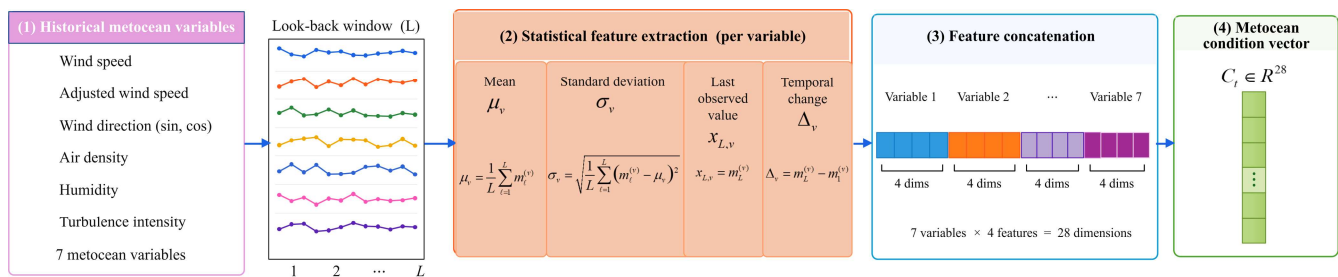


Figure 3. Construction of the metocean condition vector.

The seven metocean variables are represented as

$$M_t = [m_t^{(1)}, m_t^{(2)}, \dots, m_t^{(7)}] \in R^{L \times 7} \quad (15)$$

where $m_t^{(j)} \in R^L$ denotes the historical observations of the j -th metocean variable within the current input window.

Four statistical descriptors are extracted from each variable to characterize complementary aspects of the current condition. The window mean of the j -th variable is defined as

$$\mu_{t,j} = \frac{1}{L} \sum_{\ell=1}^L m_{t,\ell}^{(j)} \quad (16)$$

and its standard deviation is

$$\sigma_{t,j} = \sqrt{\frac{1}{L} \sum_{\ell=1}^L (m_{t,\ell}^{(j)} - \mu_{t,j})^2} \quad (17)$$

Here, $L = 96$ is the historical window length in observations (Section 3.1). This finite-window statistic involves no series truncation. The same L ensures consistent historical context across forecasting horizons.

The latest observation is defined as

$$r_{t,j} = m_{t,L}^{(j)} \quad (18)$$

while the net temporal change across the historical window is

$$\Delta_{t,j} = m_{t,L}^{(j)} - m_{t,1}^{(j)} \quad (19)$$

These four quantities characterize complementary properties of the metocean state. Specifically, $\mu_{t,j}$ represents the average operating level, $\sigma_{t,j}$ describes the local variability, $r_{t,j}$ represents the most recent state, and $\Delta_{t,j}$ describes the net temporal change over the historical window.

The local condition descriptor of the j -th metocean variable is therefore defined as

$$s_{t,j} = [\mu_{t,j}, \sigma_{t,j}, r_{t,j}, \Delta_{t,j}]^T \in R^4 \quad (20)$$

The descriptors of all seven variables are then concatenated to obtain the final metocean condition vector:

$$c_t = \text{Concat}(s_{t,1}, s_{t,2}, \dots, s_{t,7}) \in R^{28} \quad (21)$$

The resulting 28-dimensional vector summarizes each historical window from four complementary perspectives while maintaining a compact representation. Importantly, the condition vector does not replace the complete historical input sequence. The PatchTST experts continue to model detailed temporal dependencies from X_t , whereas c_t is used only by the gating network to determine the relative contribution of the experts. This separation allows the forecasting experts to focus on temporal representation learning while the gating network focuses on condition-dependent expert coordination.

2.5. Metocean-Conditioned Expert Routing and Dynamic Fusion

The metocean-conditioned expert routing and dynamic fusion mechanism is illustrated in Figure 4. Given the 28-dimensional condition vector c_t , the gating network first projects it into a latent operating-state representation and subsequently generates sample-dependent routing weights for the three PatchTST experts.

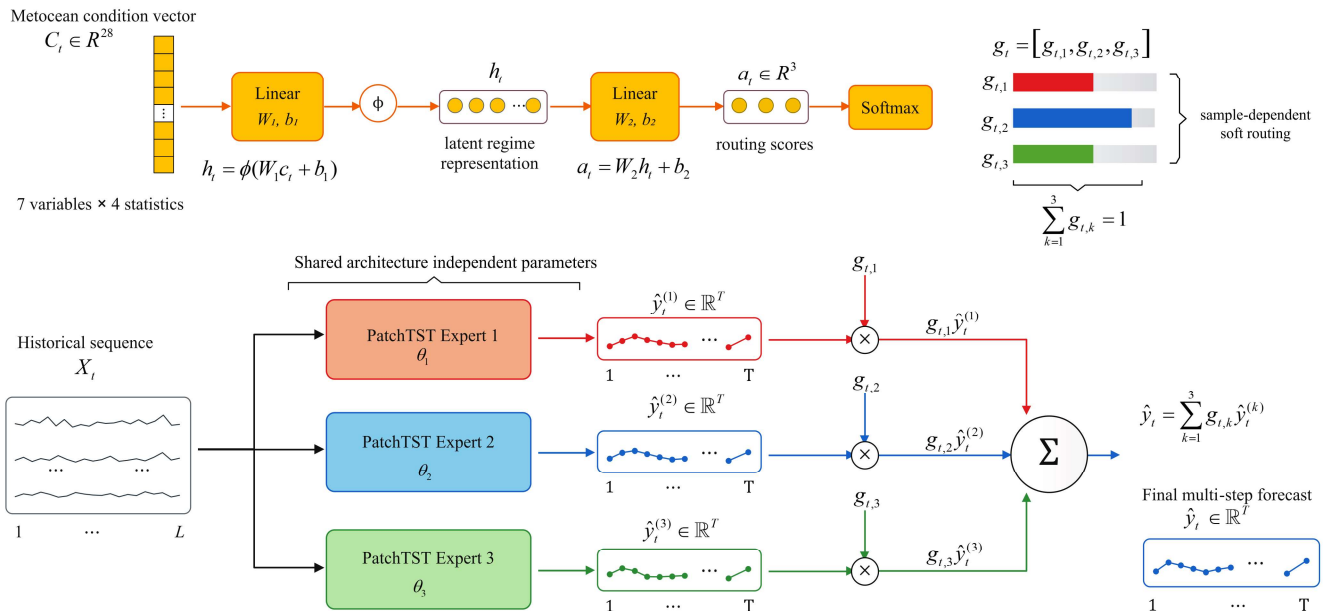


Figure 4. Architecture of the regime-aware expert routing and dynamic fusion mechanism.

In the implementation, the 28-dimensional condition vector is first normalized using LayerNorm. A two-layer multilayer perceptron (MLP), with a hidden dimension of 64, a Gaussian error linear unit (GELU) activation, and a dropout rate of 0.1, then maps this vector to three routing scores. Softmax converts these

scores into non-negative expert weights that sum to one. Each historical input window generates one weight vector, which is applied to all forecast steps when combining the three expert predictions. During training, the MSE loss of the fused forecast is backpropagated through both the gating network and the experts. This joint optimization enables the gate to learn condition-dependent expert combinations directly from forecasting errors, without predefined regime labels.

The condition vector is first projected into a latent space:

$$h_t = \phi(W_1 c_t + b_1) \quad (22)$$

where W_1 and b_1 denote the weight matrix and bias of the first mapping layer, respectively, and $\phi(\cdot)$ denotes a nonlinear activation function.

The latent representation is subsequently mapped to the expert routing scores:

$$a_t = W_2 h_t + b_2 \in R^K \quad (23)$$

The routing score associated with each expert is normalized using the softmax function:

$$g_{t,k} = \frac{\exp(a_{t,k})}{\sum_{j=1}^K \exp(a_{t,j})}, \quad k=1, \dots, K \quad (24)$$

Accordingly,

$$g_{t,k} \geq 0, \quad \sum_{k=1}^K g_{t,k} = 1 \quad (25)$$

Because the expert-weight distribution is generated from the current historical window, the contribution of each expert varies across samples. A larger $g_{t,k}$ increases the contribution of the k -th expert under the current metocean condition, whereas a smaller value reduces its influence on the fused forecast.

Rather than selecting a single expert, the proposed framework adopts soft expert fusion. All three experts produce forecasts for each input window, and the gating network continuously adjusts their relative contributions based on the current metocean conditions. Let $\hat{y}_t^{(k)}$ denote the future power sequence predicted by the k -th expert. The final multi-step forecast is obtained as

$$\hat{y}_t = \sum_{k=1}^K g_{t,k} \hat{y}_t^{(k)} \in R^H \quad (26)$$

The routing mechanism operates at the historical-window level. Thus, one expert-weight vector is generated for each input window and is applied to the complete H -step prediction produced by each expert. Different historical windows can therefore employ different fusion strategies according to their metocean conditions.

No predefined physical regime is assigned to any individual expert. Instead, the experts and the gating network are jointly optimized, allowing expert specialization and routing behavior to emerge directly from the forecasting objective. In this way, RegimeMoE-PatchTST dynamically adjusts the effective forecasting strategy according to heterogeneous offshore operating conditions without requiring manually specified regime labels.

2.6. Training Objective

RegimeMoE-PatchTST is trained in an end-to-end manner. For a mini-batch containing B samples, let $y_{b,h}$ and $\hat{y}_{b,h}$ denote the observed and predicted power values of the b -th sample at the h -th future time step, respectively.

Mean squared error is adopted as the training objective:

$$L_{\text{pred}} = \frac{1}{BH} \sum_{b=1}^B \sum_{h=1}^H (y_{b,h} - \hat{y}_{b,h})^2 \quad (27)$$

The complete parameter set is expressed as

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_k, \psi\} \quad (28)$$

where θ_k denotes the parameters of the k -th PatchTST expert and ψ denotes the parameters of the gating network.

Because the final fused prediction depends jointly on the expert outputs and the routing weights, the forecasting loss is backpropagated through both the expert branches and the gating network. Consequently, all components are optimized jointly under the same forecasting objective. No predefined regime labels are required; instead, differentiated expert representations and condition-dependent routing behavior are learned directly from forecasting errors.

3. Experimental Setup

Figure 5 summarizes the experimental workflow, including data preparation, chronological partitioning, model training, validation-based selection, and test evaluation.

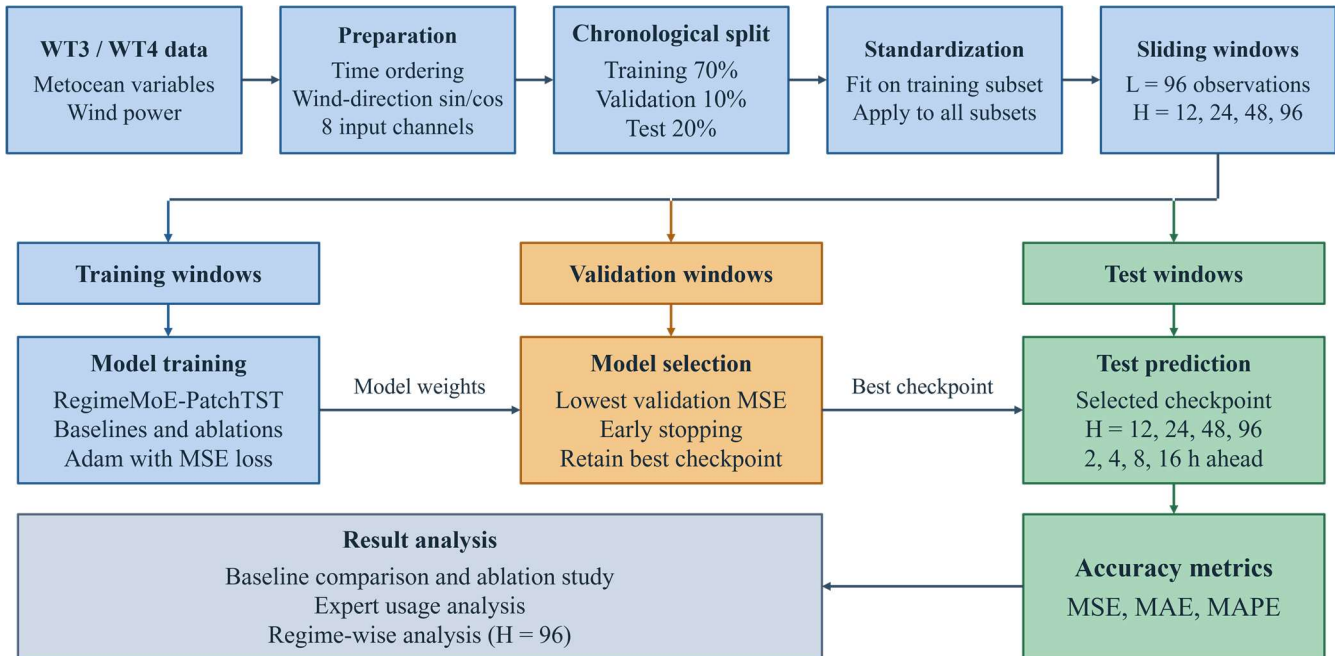


Figure 5. Experimental workflow for offshore wind power forecasting.

3.1. Datasets and Preprocessing

Two offshore wind turbine (WT) datasets, denoted as WT3 and WT4, are used to evaluate the proposed model. Both datasets are recorded at a 10-min sampling interval. WT3 contains 113,378 time steps, whereas WT4 contains 110,566 time steps. The two datasets provide simultaneous observations of the operating

conditions and power output of offshore wind turbines. Their dataset settings are summarized in Table 1. Table 2 reports descriptive statistics of the main environmental variables and normalized power.

The datasets are the offshore subset of the public Inland-Offshore Wind Farm Dataset2 [40], covering 2007–2010. The source documentation identifies a European offshore wind farm, with reported turbine specifications of approximately 3 MW rated power, 70 m hub height, and approximately 90 m rotor diameter [41]. Exact site and turbine model identifiers are not disclosed in the public documentation.

Each observation contains eight variables: wind speed V , adjusted wind speed V_{adj} , the sine and cosine components of wind direction D_{sin} and D_{cos} , air density ρ , humidity H , turbulence intensity I , and normalized wind power normPW. Since wind direction is a circular variable, directly using the angular value introduces an artificial discontinuity around the $0^\circ - 360^\circ$ boundary. The original wind direction D is therefore transformed as

$$D_{sin} = \sin\left(\frac{\pi D}{180}\right) \quad (29)$$

$$D_{cos} = \cos\left(\frac{\pi D}{180}\right) \quad (30)$$

The variables $V, V_{adj}, D_{sin}, D_{cos}, \rho, H$, and I are treated as metocean variables and are used to characterize the operating conditions within each historical input window. The normalized wind power normPW is included as a historical input and is also used as the forecasting target.

For all experiments, the historical input length is set to 96 time steps, corresponding to 16 h of past observations. Four prediction horizons are considered: 12, 24, 48, and 96 steps, corresponding to forecast lead times of 2 h, 4 h, 8 h, and 16 h, respectively. This setting allows the forecasting performance to be evaluated over both short- and relatively long-term horizons.

Table 1. Dataset settings of offshore wind turbines.

Dataset	Turbine Type	Samples	Sampling Interval	Input Length	Prediction Horizons	Input Variables	Target
WT3	Offshore wind turbine	113,378	10 min	96	12, 24, 48, 96	8 variables	normPW
WT4	Offshore wind turbine	110,566	10 min	96	12, 24, 48, 96	8 variables	normPW

Table 2. Descriptive statistics of WT3 and WT4 before standardization.

Variable	WT3	WT3	WT4	WT4
	Mean $\pm$ SD	Min–Max	Mean $\pm$ SD	Min–Max
Wind speed (m/s)	9.214 $\pm$ 3.733	1.800–24.100	9.278 $\pm$ 3.828	2.500–24.700
Air density (kg/m ³)	1.243 $\pm$ 0.024	1.170–1.331	1.242 $\pm$ 0.024	1.170–1.325
Humidity	79.239 $\pm$ 12.438	9.871–99.898	79.164 $\pm$ 12.539	9.871–99.898
Turbulence intensity	0.102 $\pm$ 0.029	0.021–0.875	0.105 $\pm$ 0.031	0.018–1.215
Normalized power	46.115 $\pm$ 33.637	0.003–100.063	46.909 $\pm$ 33.839	0.003–100.043

Statistics use all available observations. SD denotes sample standard deviation. Humidity and normalized power retain their source-file scales.

3.2. Experimental Settings

To prevent information leakage from future observations, WT3 and WT4 are chronologically divided into training, validation, and test sets. The first 70% of the samples are used for training, the subsequent 10% for model selection and early stopping, and the remaining 20% for final evaluation. All models use

the same data partitions and evaluation procedure to ensure a fair comparison. Model selection is performed exclusively on the validation set, and the test set is not involved in either training or hyperparameter tuning.

The proposed RegimeMoE-PatchTST consists of three PatchTST experts with identical architectures. The regime gate receives a 28-dimensional metocean condition vector, constructed from the mean, standard deviation, last observed value, and temporal change of the seven metocean variables within the historical window. Based on this condition representation, the gate generates sample-dependent expert weights and dynamically combines the predictions of the three experts. To isolate the contribution of the proposed mechanism, RegimeMoE-PatchTST and PatchTST use the same backbone configuration.

All models are trained using the Adam optimizer with a batch size of 128 and an initial learning rate of 1×10^{-4} . The maximum number of training epochs is set to 50, and early stopping with a patience of 3 epochs is applied to mitigate overfitting. The model parameters associated with the lowest validation loss are retained for evaluation on the test set.

3.3. Baseline Models and Evaluation Metrics

To evaluate the proposed model across different forecasting architectures, six representative methods are selected as baselines: FEDformer [22], TimesNet [24], iTransformer [26], TiDE [25], PAttn [42], and PatchTST [27]. These methods cover frequency-domain modeling, multiscale periodic pattern extraction, cross-variable dependency learning, multilayer-perceptron-based forecasting, and patch-based temporal representation, providing a broad comparison across different time-series forecasting paradigms.

Forecasting performance is evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), which are defined as

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (31)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (32)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (33)$$

where N denotes the total number of evaluated predictions, and y_i and $\hat{y}_i$ denote the corresponding observed and predicted values, respectively. MSE places greater emphasis on large forecasting deviations, whereas MAE measures the average absolute magnitude of the prediction error. Therefore, MSE and MAE are adopted as the primary evaluation metrics. MAPE is additionally reported to characterize the relative forecasting error under different power levels. Lower values of all three metrics indicate better forecasting performance.

4. Results and Discussion

4.1. Main Forecasting Results

Tables 3 and 4 report the forecasting performance of the evaluated models on WT3 and WT4, respectively. The best result in each column is highlighted in bold, while the second-best result is underlined. Four forecasting horizons are considered, namely 12, 24, 48, and 96 steps, corresponding to 2, 4, 8, and 16 h ahead forecasting. As expected, the forecasting errors generally increase with the prediction horizon, reflecting the growing uncertainty associated with longer-range offshore wind power forecasting. Despite this increase, RegimeMoE-PatchTST achieves the lowest MSE on both turbines at all four horizons, indicating that the proposed framework provides consistent improvements across different forecasting ranges.

Table 3. Forecasting performance comparison of different models on WT3.

Model	12 MSE	12 MAE	12 MAPE	24 MSE	24 MAE	24 MAPE	48 MSE	48 MAE	48 MAPE	96 MSE	96 MAE	96 MAPE
FEDformer	0.1852	0.2887	<u>1.6472</u>	0.2987	0.3763	<u>1.9122</u>	0.4862	0.5107	<u>2.1830</u>	<u>0.7218</u>	0.6526	2.4538
TimesNet	0.1918	0.2854	1.6601	0.3261	0.3894	2.0900	0.5298	0.5107	2.4309	0.8300	0.6625	2.7794
iTransformer	0.1890	0.2750	1.8287	0.3235	0.3764	2.0722	0.5484	0.5264	2.4737	0.8091	0.6589	2.8499
TiDE	0.1890	0.2725	1.6756	0.3007	0.3753	2.0102	0.5050	0.5272	2.4042	0.7744	0.6866	2.7585
PAttn	<u>0.1769</u>	0.2637	1.6823	<u>0.2970</u>	0.3635	2.0038	<u>0.4843</u>	0.4979	2.2467	0.7321	<u>0.6450</u>	2.3905
PatchTST	0.1801	0.2734	1.6482	0.2971	0.3693	1.9542	0.4856	0.5022	2.2433	0.7354	0.6494	<u>2.3552</u>
Proposed	0.1729	<u>0.2647</u>	1.6181	0.2879	<u>0.3641</u>	1.8808	0.4736	<u>0.5010</u>	2.1280	0.7026	0.6428	2.1537

Note: Bold values indicate the best result; underlined values indicate the second-best result in each column.

Table 4. Forecasting performance comparison of different models on WT4.

Model	12 MSE	12 MAE	12 MAPE	24 MSE	24 MAE	24 MAPE	48 MSE	48 MAE	48 MAPE	96 MSE	96 MAE	96 MAPE
FEDformer	0.1742	0.2776	<u>1.2805</u>	<u>0.2850</u>	0.3633	1.6538	<u>0.4710</u>	0.4983	<u>1.8868</u>	<u>0.7032</u>	0.6369	2.1510
TimesNet	0.1890	0.2856	1.4556	0.3153	0.3802	1.7046	0.5262	0.5072	2.1001	0.8041	0.6559	2.4817
iTransformer	0.1881	0.2875	1.3705	0.3104	0.3759	1.7570	0.5228	0.5063	2.2069	0.7737	0.6442	2.4644
TiDE	0.1718	0.2676	1.3245	0.2914	0.3789	1.7403	0.4937	0.5225	2.0723	0.7705	0.6494	2.3976
PAttn	<u>0.1696</u>	<u>0.2602</u>	1.2836	0.2859	<u>0.3567</u>	1.6962	0.4747	0.4891	1.9841	0.7118	<u>0.6351</u>	<u>2.1094</u>
PatchTST	0.1722	0.2684	1.3098	0.2870	0.3651	<u>1.6510</u>	0.4735	<u>0.4940</u>	1.9438	0.7102	0.6358	2.1242
Proposed	0.1649	0.2597	1.2672	0.2747	0.3557	1.5998	0.4592	0.4944	1.8135	0.6817	0.6289	2.0516

Note: Bold values indicate the best result; underlined values indicate the second-best result in each column.

On WT3, RegimeMoE-PatchTST obtains MSE values of 0.1729, 0.2879, 0.4736, and 0.7026 at the 12-, 24-, 48-, and 96-step horizons, respectively. Compared with the PatchTST backbone, these results correspond to MSE reductions of 4.00%, 3.08%, 2.47%, and 4.46%, respectively. The improvement is therefore observed across all forecasting horizons rather than concentrated within a particular prediction range. In terms of MAE, the proposed model also consistently outperforms PatchTST, reducing the error from 0.2734 to 0.2647, from 0.3693 to 0.3641, from 0.5022 to 0.5010, and from 0.6494 to 0.6428. Although PAttn achieves slightly lower MAE values at the 12-, 24-, and 48-step horizons, RegimeMoE-PatchTST remains close to the best-performing method and obtains the lowest MAE at the 96-step horizon. Moreover, the proposed model achieves the lowest MAPE at all four horizons, further demonstrating its stable forecasting performance across different power levels.

A similar pattern can be observed on WT4. RegimeMoE-PatchTST achieves MSE values of 0.1649, 0.2747, 0.4592, and 0.6817 at the four forecasting horizons, corresponding to reductions of 4.21%, 4.27%, 3.03%, and 4.02% relative to PatchTST, respectively. The proposed model also achieves the lowest MAE at the 12-, 24-, and 96-step horizons. At the 48-step horizon, its MAE is 0.4944, which remains close to the PatchTST result of 0.4940, although PAttn obtains the lowest value of 0.4891 in this setting. Meanwhile, RegimeMoE-PatchTST again produces the lowest MAPE at all four horizons. These results show that the performance gain of the proposed model is reproducible across the two offshore turbines and is not restricted to a specific forecasting horizon.

The advantage of RegimeMoE-PatchTST is also evident when it is compared with the strongest competing baseline in each turbine–horizon setting rather than only with its PatchTST backbone. Based on the MSE results in Tables 3 and 4, the proposed model further reduces the MSE by approximately 2.21–3.61% relative to the best non-proposed result in the corresponding setting, with an average reduction of approximately 2.77% across all eight turbine–horizon combinations. This observation is important because it indicates that the improvement cannot be explained merely as an incremental enhancement over

PatchTST; instead, the proposed framework remains advantageous when compared with a broader set of competitive forecasting architectures.

Figure 6 further illustrates the variations in MSE and MAE as the prediction horizon increases. For all models, both metrics generally increase as the horizon extends from 12 to 96 steps. Nevertheless, RegimeMoE-PatchTST maintains the lowest MSE trajectory on both WT3 and WT4 throughout the evaluated horizons. It also remains among the best-performing methods in terms of MAE, with only small differences from the lowest MAE in the few settings where another baseline performs slightly better. Overall, the results demonstrate that the proposed metocean-conditioned mixture-of-experts mechanism provides consistent MSE improvements over different turbines and forecasting horizons while maintaining competitive performance in the complementary error metrics. This suggests that dynamically adapting expert contributions to the metocean conditions in each historical window can improve the robustness of the PatchTST forecasting backbone across heterogeneous offshore operating conditions.

Across all eight turbine–horizon settings, the proposed model reduces MSE relative to PatchTST, with a mean relative reduction of 3.69%. For planning 2–16 h ahead, more accurate wind-power forecasts can inform conventional-generation scheduling and reserve allocation by improving estimates of the wind supply available to meet demand [1,5]. The larger reductions of 7.7% and 8.1% under high-wind and high-turbulence conditions in the 16-h analysis (Section 4.4) indicate greater accuracy gains in these challenging operating regimes.

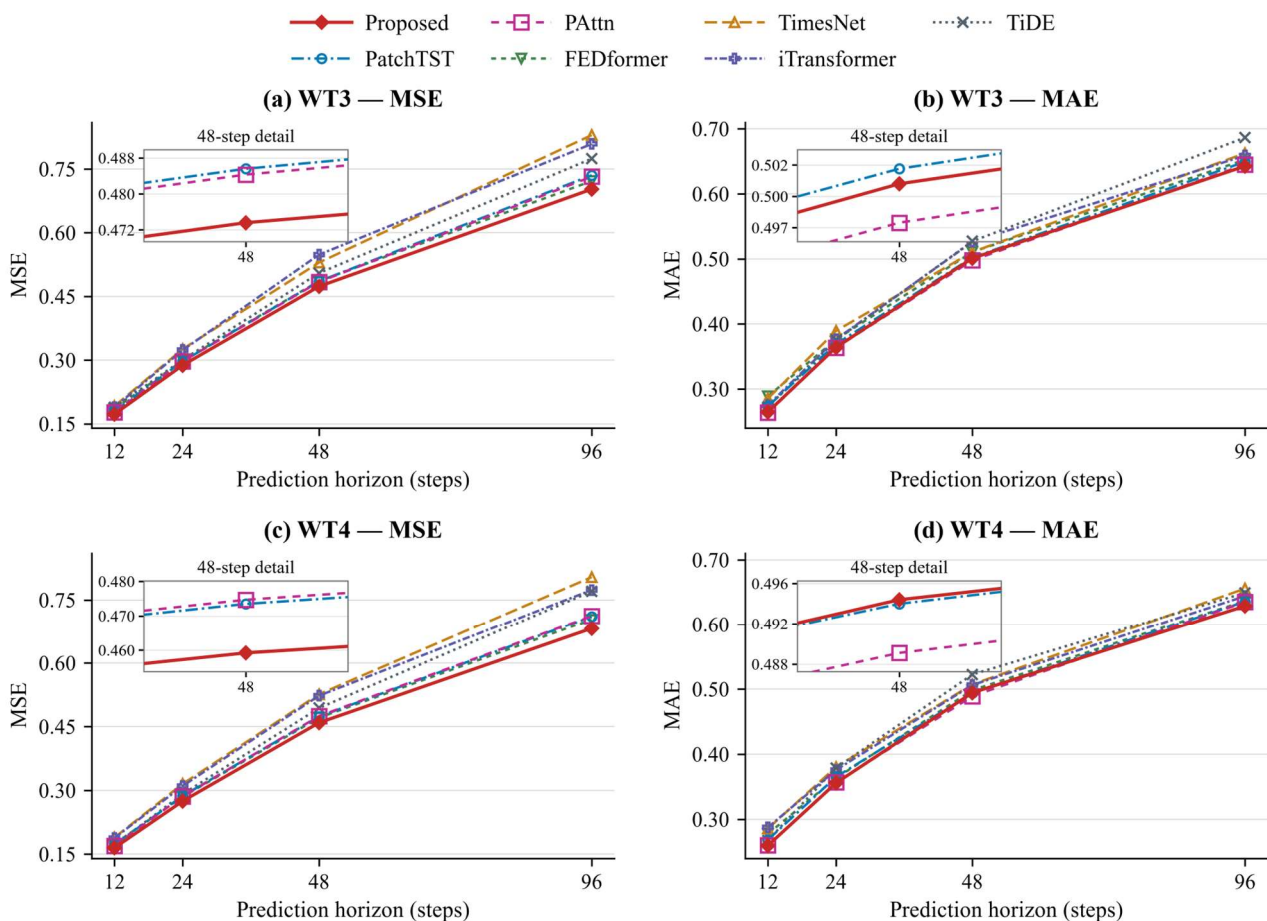


Figure 6. Variations in MSE and MAE with the prediction horizon for different forecasting models on WT3 and WT4. (a) WT3 MSE; (b) WT3 MAE; (c) WT4 MSE; (d) WT4 MAE. Insets magnify the 48-step region for Proposed, PatchTST, and PAttn.

4.2. Ablation Study

To investigate the contribution of the proposed components and determine whether the performance improvement can be attributed merely to the increased model capacity introduced by multiple experts, an ablation study is conducted on WT3 and WT4. Four model variants are compared. PatchTST represents the single-expert backbone. The w/o Gate variant retains three independently parameterized PatchTST experts but removes the dynamic routing mechanism, and the expert predictions are combined by simple averaging. The w/o Metocean variant retains the gating structure but removes the input-dependent metocean condition information. The Proposed model represents the complete RegimeMoE-PatchTST framework, including both metocean-conditioned routing and dynamic expert fusion. The results over the four forecasting horizons are reported in Table 5.

Table 5. Ablation results of RegimeMoE-PatchTST on WT3 and WT4.

Dataset	Horizon (Steps)	PatchTST MSE	PatchTST MAE	w/o Gate MSE	w/o Gate MAE	w/o Metocean MSE	w/o Metocean MAE	Proposed MSE	Proposed MAE
WT3	12	0.1801	0.2734	0.1770	0.2640	0.1763	0.2626	0.1729	0.2647
	24	0.2971	0.3693	0.2946	0.3620	0.2948	0.3638	0.2879	0.3641
	48	0.4856	0.5022	0.4913	0.5017	0.4886	0.4980	0.4736	0.5010
	96	0.7354	0.6494	0.7431	0.6510	0.7348	0.6477	0.7026	0.6428
WT4	12	0.1722	0.2684	0.1694	0.2607	0.1695	0.2606	0.1649	0.2597
	24	0.2870	0.3651	0.2843	0.3570	0.2879	0.3594	0.2747	0.3557
	48	0.4735	0.4940	0.4790	0.4941	0.4790	0.4986	0.4592	0.4944
	96	0.7102	0.6358	0.7149	0.6379	0.7122	0.6357	0.6817	0.6289

In terms of MSE, the complete model achieves the best performance in all eight turbine–horizon combinations. In contrast, simply introducing multiple experts without dynamic routing does not provide consistent improvements. The w/o Gate variant performs better than PatchTST in several short-horizon settings, such as the 12- and 24-step forecasts on both turbines, but its advantage disappears as the forecasting horizon increases. For example, on WT3 at the 96-step horizon, the MSE of w/o Gate increases to 0.7431, compared with 0.7354 for PatchTST, whereas the complete model reduces the MSE to 0.7026. Similar degradation is observed on WT4 at the 48- and 96-step horizons, where simple expert averaging yields MSE values of 0.4790 and 0.7149, respectively, both higher than those of the PatchTST backbone. These results indicate that the gain of RegimeMoE-PatchTST cannot be explained solely by the use of additional experts or increased parameter capacity. Instead, dynamically adjusting the contributions of different experts is more effective than assigning equal weights to them.

Removing the metocean condition information also weakens the effectiveness of the mixture-of-experts structure. At the 96-step horizon on WT3, the w/o Metocean variant obtains an MSE of 0.7348, which is nearly identical to the PatchTST result of 0.7354, while the complete model achieves a substantially lower MSE of 0.7026. A similar pattern occurs on WT4 at the 24-step horizon: w/o Metocean produces an MSE of 0.2879, slightly worse than the 0.2870 obtained by PatchTST, whereas the complete model reduces the error to 0.2747. Across all eight settings, RegimeMoE-PatchTST consistently outperforms the w/o Metocean variant, indicating that a gating structure without input-dependent metocean information is insufficient to achieve the same level of forecasting performance. The window-level metocean representation, therefore, provides useful condition information for sample-dependent expert routing.

The differences in MAE are smaller than those observed for MSE. Nevertheless, the complete model achieves the lowest MAE at the 96-step horizon on WT3 and at the 12-, 24-, and 96-step horizons on WT4, while remaining close to the best-performing variant in the other settings. For instance, at the 48-step horizon on WT4, the MAE of the complete model is 0.4944, only 0.0004 higher than that of PatchTST.

More importantly, RegimeMoE-PatchTST is the only variant that consistently reduces MSE relative to PatchTST across both turbines and all forecasting horizons. Across the eight turbine–horizon combinations, the proposed model achieves an average relative MSE reduction of approximately 3.56% over w/o Gate and 3.43% over w/o Metocean. These results provide consistent evidence that the forecasting gains depend on the joint use of metocean condition modeling and dynamic expert routing, rather than on additional model capacity alone.

4.3. Expert Routing Behavior Analysis

To further examine how the dynamic gate utilizes different experts, the expert winning ratio is analyzed for each forecasting horizon. For each test sample, the expert with the largest gating weight is considered the winning expert, and its winning ratio is defined as the proportion of test samples for which that expert receives the highest weight. This metric provides a simple view of the dominant routing decisions made by the gate across the test set. Figure 7 presents the winning ratios of the three experts on WT3 and WT4 under different forecasting horizons.

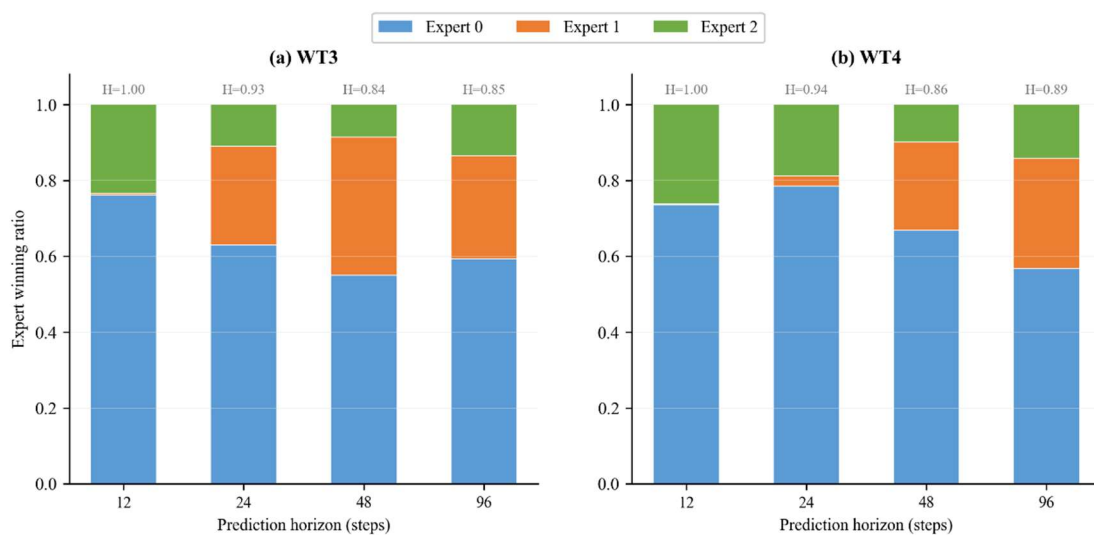


Figure 7. Expert winning ratios of RegimeMoE-PatchTST under different forecasting horizons. (a) WT3; (b) WT4.

As shown in Figure 7, the routing decisions are not persistently concentrated on a single expert across all forecasting settings. At the 12-step horizon, Expert 0 receives the highest weight for 76.1% and 73.5% of the test samples on WT3 and WT4, respectively, indicating that it plays a dominant role in short-horizon forecasting. However, the routing distribution becomes more diversified at longer horizons. For example, the winning ratio of Expert 1 reaches 36.4% at the 48-step horizon on WT3 and 29.1% at the 96-step horizon on WT4. Expert 2 also retains a non-negligible proportion of winning samples across the evaluated settings. These observations indicate that although one expert may dominate under a particular forecasting configuration, the gate does not consistently collapse to the same expert across different tasks.

The expert winning ratios also do not change monotonically with the forecasting horizon. This suggests that the routing behavior is not determined solely by the prediction horizon. Since the gating weights are generated from the metocean condition representation of each historical window, different samples within the same forecasting task can favor different experts. The broadly similar expert-utilization patterns observed on WT3 and WT4 further suggest that the learned routing behavior is not restricted to a single turbine.

Overall, the expert-usage analysis provides evidence that the proposed gate learns differentiated routing patterns rather than relying on a fixed expert or a uniform expert allocation. Combined with the ablation results in Section 4.2, these observations support the role of dynamic expert routing in the proposed

framework. It should be noted, however, that the winning ratio characterizes only the frequency with which each expert receives the largest gating weight; therefore, it is used here to analyze expert utilization and routing diversity rather than to assign explicit physical meanings to individual experts.

4.4. Metocean-Regime-Wise Analysis

To further evaluate the adaptability of RegimeMoE-PatchTST under heterogeneous offshore operating conditions, a metocean-regime-wise analysis is conducted for the 96-step forecasting task. This setting corresponds to 16 h ahead forecasting and represents the longest and most challenging prediction horizon considered in this study. The test samples are grouped according to the wind-speed level, turbulence intensity, and wind-direction variation within each historical input window. The MSE values of PatchTST and RegimeMoE-PatchTST are then calculated separately for each regime. To reduce the influence of turbine-specific operating characteristics and to highlight the overall trend, Figure 8 presents regime-wise results averaged over WT3 and WT4. The number of test samples in each regime is also provided to indicate the scale of each comparison.

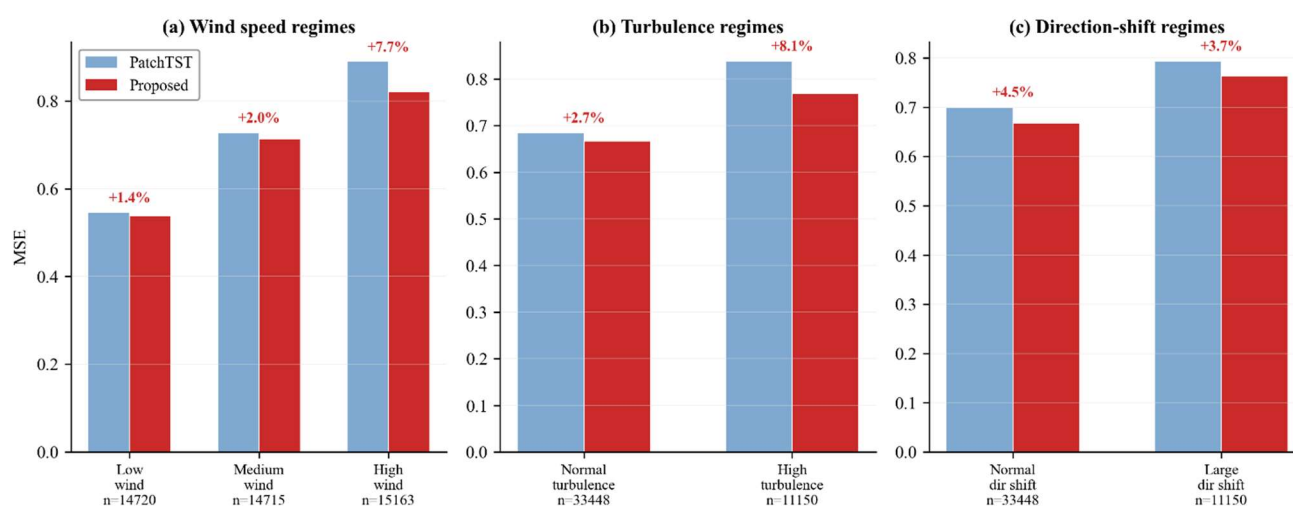


Figure 8. Regime-wise MSE comparison between PatchTST and the proposed model under different metocean conditions. (a) Wind-speed regimes; (b) turbulence regimes; (c) wind-direction-shift regimes. Results are averaged over WT3 and WT4; n denotes the number of test samples.

For the wind-speed analysis, the test samples are divided into low-, medium-, and high-wind regimes based on the mean wind speed over the historical input window. For turbulence intensity, the samples are grouped into normal- and high-turbulence regimes. Wind-direction conditions are further divided into normal and large direction-shift regimes according to the magnitude of directional variation within the input window.

As shown in Figure 8, RegimeMoE-PatchTST achieves lower MSE than PatchTST in all seven metocean regimes, indicating that the forecasting improvement is not restricted to a particular environmental condition. Under the low- and medium-wind regimes, the proposed model reduces MSE by 1.4% and 2.0%, respectively. The improvement becomes substantially larger under the high-wind regime, where the MSE reduction reaches 7.7%. This result suggests that the proposed framework provides greater benefits under operating conditions in which the wind-to-power relationship becomes more difficult to model with a single shared forecasting function.

A similar tendency is observed for the turbulence-based groups. Compared with PatchTST, RegimeMoE-PatchTST reduces MSE by 2.7% under normal-turbulence conditions and by 8.1% under high-turbulence conditions. The considerably larger improvement in the high-turbulence regime indicates that the proposed dynamic expert fusion is particularly beneficial when the historical sequence exhibits

stronger short-term variability. Under such conditions, relying on a single forecasting representation may be less effective, whereas condition-dependent combinations of multiple experts provide additional flexibility for modeling heterogeneous temporal dynamics.

For the wind-direction groups, the proposed model achieves MSE reductions of 4.5% and 3.7% under the normal and large direction-shift regimes, respectively. Although the improvement under large direction shifts is less pronounced than those observed under high wind speed and high turbulence, RegimeMoE-PatchTST still consistently outperforms PatchTST. This result indicates that the proposed framework retains its forecasting advantage even when the directional state of the wind varies substantially within the historical window.

Overall, the regime-wise results show that the advantage of RegimeMoE-PatchTST is maintained across a diverse range of metocean conditions, while the largest gains are observed under high-wind and high-turbulence regimes. This pattern is consistent with the motivation of the proposed method: when offshore operating conditions become more heterogeneous, and the relationship between historical observations and future power becomes more difficult to represent using a single shared forecasting function, metocean-conditioned expert routing can provide a more flexible forecasting strategy. It should be emphasized that this analysis evaluates forecasting performance under predefined metocean regimes; it does not by itself establish a one-to-one correspondence between individual experts and specific physical regimes.

4.5. Computational Cost and Resource Constraints

To assess computational cost, the saved WT3 checkpoints of both models were benchmarked at all four forecasting horizons on an NVIDIA GeForce RTX 5060 Ti graphics processing unit (GPU; 16 GB), using PyTorch 2.8.0 and CUDA 12.8. Both models used identical inputs and single-precision inference at batch sizes of 1 and 32. Table 6 reports parameter counts, inference latency, and peak allocated GPU memory.

Table 6. Supplementary inference cost comparison using the saved WT3 checkpoints.

Horizon (Steps)	Model	Parameters	Latency, B = 1 (ms)	Latency, B = 32 (ms)	Peak GPU Memory, B = 32 (MiB)
12	PatchTST	285,708	1.09 [1.03–1.36]	1.11 [1.07–1.43]	25.41
12	RegimeMoE-PatchTST	859,231	3.21 [3.04–3.82]	3.31 [3.05–3.88]	32.22
24	PatchTST	304,152	1.05 [1.02–1.32]	1.11 [1.09–1.40]	25.48
24	RegimeMoE-PatchTST	914,563	3.34 [3.05–3.96]	3.39 [3.09–3.91]	32.46
48	PatchTST	341,040	1.04 [1.01–1.39]	1.15 [1.12–1.45]	25.62
48	RegimeMoE-PatchTST	1,025,227	3.34 [3.07–3.89]	3.45 [3.14–3.80]	32.92
96	PatchTST	414,816	1.04 [1.01–1.20]	1.12 [1.07–1.61]	25.90
96	RegimeMoE-PatchTST	1,246,555	3.21 [3.04–3.90]	3.28 [3.07–3.69]	33.86

B denotes batch size. Per-batch latency is the median [25th–75th percentiles] of 500 calls (five rounds; 20 warm-up calls per round), with GPU synchronization before and after each call; disk I/O, external preprocessing, and data transfer are excluded. Peak memory at B = 32 is the PyTorch allocator-tracked allocation, including library workspace but excluding unused cache and the CUDA context. MiB = 2^{20} bytes.

RegimeMoE-PatchTST has approximately three times as many parameters as PatchTST and requires 2.95–3.20 times its single-window inference latency. Thus, the forecasting gains entail additional computational cost because all three experts are evaluated for every input. These measurements describe model inference on the tested hardware, excluding training costs and end-to-end deployment latency. Deployment under other conditions requires a separate assessment. Parameter sharing, sparse routing, and lightweight experts may reduce this burden.

5. Conclusions

This study proposes RegimeMoE-PatchTST, a metocean-conditioned mixture-of-experts framework for multi-step offshore wind power forecasting under heterogeneous operating conditions. Instead of assigning historical samples to predefined discrete regimes, the proposed method constructs a compact window-level metocean condition representation from seven environmental variables and uses it to generate sample-dependent routing weights for three independently parameterized PatchTST experts. The complete multi-step forecasts produced by the experts are then dynamically combined through a soft gating mechanism, enabling the model to adapt the contributions of different experts according to the environmental state of each historical window.

Experiments were conducted on two offshore wind turbine datasets, WT3 and WT4, with prediction horizons of 12, 24, 48, and 96 steps, corresponding to 2, 4, 8, and 16 h ahead forecasting. RegimeMoE-PatchTST achieved the lowest MSE in all eight turbine–horizon combinations and reduced the average MSE by approximately 3.69% relative to the PatchTST backbone, while maintaining competitive MAE performance across the evaluated settings. These results indicate that the proposed framework provides consistent forecasting improvements across different turbines and prediction horizons rather than being effective only under a particular experimental configuration.

The ablation results further show that the performance improvement cannot be attributed solely to the increased model capacity introduced by multiple experts. Simply averaging the predictions of three PatchTST experts does not provide consistent gains, while removing the input-dependent metocean condition information also weakens the effectiveness of the gating mechanism. These observations provide consistent evidence that the forecasting gains depend on the joint use of metocean condition modeling and dynamic expert routing. In addition, the expert routing behavior analysis shows differentiated expert-utilization patterns across forecasting settings without persistent expert collapse. The metocean-regime-wise evaluation further demonstrates that the proposed model maintains lower forecasting errors across all evaluated environmental regimes, with MSE reductions of 7.7% and 8.1% under high-wind and high-turbulence conditions, respectively. These results suggest that metocean-conditioned expert routing is particularly beneficial under challenging offshore operating conditions, especially high-wind and high-turbulence regimes.

Several limitations remain. First, the experiments are conducted on only two offshore wind turbines, and the generalization capability of the proposed framework across different wind farms, turbine types, and climatic regions requires further investigation. Second, employing multiple complete PatchTST experts increases the parameter count and computational cost compared with a single backbone model. Future work may therefore explore sparse routing, parameter sharing, and lightweight expert architectures to improve computational efficiency. In addition, the present study focuses on deterministic multi-step forecasting and does not explicitly quantify predictive uncertainty. Future research will investigate cross-farm transfer, probabilistic forecasting, physics-informed constraints, and the integration of richer marine environmental information to further improve the generalization, efficiency, interpretability, and practical applicability of metocean-conditioned expert forecasting.

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 (OpenAI) to improve its language and readability. 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

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

Ethics Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

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

Funding

This study is 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 that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wang Y, Zou R, Liu F, Zhang L, Liu Q. A review of wind speed and wind power forecasting with deep neural networks. *Appl. Energy* **2021**, 304, 117766. DOI:10.1016/j.apenergy.2021.117766
2. Deng X, Shao H, Hu C, Jiang D, Jiang Y. Wind power forecasting methods based on deep learning: A survey. *Comput. Model. Eng. Sci.* **2020**, 122, 273–301. DOI:10.32604/cmes.2020.08768
3. Liu H, Chen C. Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Appl. Energy* **2019**, 249, 392–408. DOI:10.1016/j.apenergy.2019.04.188
4. Wang Y, Hu Q, Li L, Foley AM, Srinivasan D. Approaches to wind power curve modeling: A review and discussion. *Renew. Sustain. Energy Rev.* **2019**, 116, 109422. DOI:10.1016/j.rser.2019.109422
5. Tawn R, Browell J. A review of very short-term wind and solar power forecasting. *Renew. Sustain. Energy Rev.* **2022**, 153, 111758. DOI:10.1016/j.rser.2021.111758
6. Lin Z, Liu X. Wind power forecasting of an offshore wind turbine based on high-frequency SCADA data and deep learning neural network. *Energy* **2020**, 201, 117693. DOI:10.1016/j.energy.2020.117693
7. Hanifi S, Zare-Behtash H, Cammarano A, Lotfian S. Offshore wind power forecasting based on WPD and optimised deep learning methods. *Renew. Energy* **2023**, 218, 119241. DOI:10.1016/j.renene.2023.119241
8. 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
9. 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
10. Ma Z, Mei G. A hybrid attention-based deep learning approach for wind power prediction. *Appl. Energy* **2022**, 323, 119608. DOI:10.1016/j.apenergy.2022.119608
11. Liu C, Zhang X, Mei S, Zhen Z, Jia M, Li Z, et al. Numerical weather prediction enhanced wind power forecasting: Rank ensemble and probabilistic fluctuation awareness. *Appl. Energy* **2022**, 313, 118769. DOI:10.1016/j.apenergy.2022.118769
12. Liu C, Zhang X, Mei S, Zhou Q, Fan H. Series-wise attention network for wind power forecasting considering temporal lag of numerical weather prediction. *Appl. Energy* **2023**, 336, 120815. DOI:10.1016/j.apenergy.2023.120815

13. Hu Y, Liu H, Wu S, Zhao Y, Wang Z, Liu X. Temporal collaborative attention for wind power forecasting. *Appl. Energy* **2024**, 357, 122502. DOI:10.1016/j.apenergy.2023.122502
14. Dai X, Liu GP, Hu W. An online-learning-enabled self-attention-based model for ultra-short-term wind power forecasting. *Energy* **2023**, 272, 127173. DOI:10.1016/j.energy.2023.127173
15. Tang Y, Yang K, Zheng Y, Ma L, Zhang S, Zhang Z. Wind power forecasting: A transfer learning approach incorporating temporal convolution and adversarial training. *Renew. Energy* **2024**, 224, 120200. DOI:10.1016/j.renene.2024.120200
16. Hu J, Hu W, Cao D, Huang Y, Chen J, Li Y, et al. Bayesian averaging-enabled transfer learning method for probabilistic wind power forecasting of newly built wind farms. *Appl. Energy* **2024**, 355, 122185. DOI:10.1016/j.apenergy.2023.122185
17. Wang S, Zhang W, Sun Y, Trivedi A, Chung CY, Srinivasan D. Wind power forecasting in the presence of data scarcity: A very short-term conditional probabilistic modeling framework. *Energy* **2024**, 291, 130305. DOI:10.1016/j.energy.2024.130305
18. de Azevedo Takara L, Teixeira AC, Yazdanpanah H, Mariani VC, dos Santos Coelho L. Optimizing multi-step wind power forecasting: Integrating advanced deep neural networks with stacking-based probabilistic learning. *Appl. Energy* **2024**, 369, 123487. DOI:10.1016/j.apenergy.2024.123487
19. Lim B, Arık SÖ, Loeff N, Pfister T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, 37, 1748–1764. DOI:10.1016/j.ijforecast.2021.03.012
20. 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
21. 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. DOI:10.48550/arXiv.2106.13008
22. 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. DOI:10.48550/arXiv.2201.12740
23. Zeng A, Chen M, Zhang L, Xu Q. Are Transformers effective for time series forecasting? *Proc. AAAI Conf. Artif. Intell.* **2023**, 37, 11121–11128. DOI:10.1609/aaai.v37i9.26317
24. 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 International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. DOI:10.48550/arXiv.2210.02186
25. Das A, Kong W, Leach A, Mathur SK, Sen R, Yu R. Long-term forecasting with TiDE: Time-series dense encoder. *Trans. Mach. Learn. Res.* **2023**. Available online: <https://openreview.net/forum?id=pCbC3aQB5W> (accessed on 29 September 2026).
26. 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 International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024. DOI:10.48550/arXiv.2310.06625
27. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: Long-term forecasting with Transformers. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. DOI:10.48550/arXiv.2211.14730
28. Wang S, Sun Y, Zhang W, Chung CY, Srinivasan D. Very short-term wind power forecasting considering static data: An improved transformer model. *Energy* **2024**, 312, 133577. DOI:10.1016/j.energy.2024.133577
29. Xu S, Wang Y, Xu X, Shi G, Huang H, Zheng Y. A PatchTST-GRU based heterogeneous seq2seq model with numerical weather prediction refinement for multi-step wind power forecasting. *Sci. Rep.* **2025**, 15, 16547. DOI:10.1038/s41598-025-00741-9
30. Qiu H, Shi K, Wang R, Zhang L, Liu X, Cheng X. A novel temporal–spatial graph neural network for wind power forecasting considering blockage effects. *Renew. Energy* **2024**, 227, 120499. DOI:10.1016/j.renene.2024.120499
31. Gao J, Cheng Y, Zhang D, Chen Y. Physics-constrained wind power forecasting aligned with probability distributions for noise-resilient deep learning. *Appl. Energy* **2025**, 383, 125295. DOI:10.1016/j.apenergy.2025.125295
32. Cai C, Shi Q, Jin Y, Hua M, Tao Y, 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
33. Isik M, Yalcinkaya MA. Integrated regime-aware wind power forecasting using multi-altitude meteorological features and hybrid machine learning. *Front. Energy Res.* **2026**, 13, 1686125. DOI:10.3389/fenrg.2025.1686125
34. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. *Neural Comput.* **1991**, 3, 79–87. DOI:10.1162/neco.1991.3.1.79

35. Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le QV, Hinton GE, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
36. Shi X, Wang S, Nie Y, Li D, Ye Z, Wen Q, et al. Time-MoE: Billion-scale time series foundation models with mixture of experts. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025.
37. Liu X, Liu J, Woo G, Aksu T, Liang Y, Zimmermann R, et al. Moirai-MoE: Empowering time series foundation models with sparse mixture of experts. In Proceedings of the 42nd International Conference on Machine Learning, Vancouver, BC, Canada, 13–19 July 2025; Volume 267, pp. 38940–38962.
38. Hossain MT, Cai J, Han J, Han X, Li Z. Mathematically-guided ultra-short-term wind power forecasting based on adaptive mixture of experts. *Energy AI* **2026**, *24*, 100742. DOI:10.1016/j.egyai.2026.100742
39. Zhao C, Wu Y, Pan J, Chen P. A physics-aware dynamic graph and mixture-of-experts framework for wind power forecasting. *Int. J. Electr. Power Energy Syst.* **2026**, *176*, 111769. DOI:10.1016/j.ijepes.2026.111769
40. Ding Y. Inland-Offshore Wind Farm Dataset2, Version v1. Zenodo. 2021. Available online: <https://zenodo.org/record/5516554> (accessed on 16 October 2026).
41. Ding Y. *Data Science for Wind Energy*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2019. DOI:10.1201/9780429490972
42. Tan M, Merrill MA, Gupta V, Althoff T, Hartvigsen T. Are language models actually useful for time series forecasting? *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 60162–60191. DOI:10.52202/079017-1922