MS-ECFormer: A Multi-Scale Endogenous–Exogenous Coupling Transformer for Wind Power Forecasting

1st Yuchi Liu
School of Computer Science (School of Artificial Intelligence)
South-Central Minzu University
Wuhan, Chain
2023110237@mail.scuec.edu.cn

2nd Zheng Ye*
School of Artificial Intelligence and Security
Cyberspace Security University of China
Wuhan, China
yezh0716@gmail.com

***Abstract*—Accurate wind power forecasting is critical for the reliable integration and operation of renewable energy systems, yet it remains challenging due to strong non-stationarity in wind dynamics and complex interactions between historical power outputs and meteorological variables. Wind power generation exhibits multi-scale temporal patterns ranging from short-term fluctuations to long-term trends. However, existing deep learning–based forecasting models often struggle to capture such multi-scale dependencies while effectively modeling the coupling between endogenous variables (historical power) and exogenous variables (meteorological factors).To address these challenges, we propose MS-ECFormer, a Multi-Scale Endogenous–Exogenous Coupling Transformer for wind power forecasting. MS-ECFormer learns multi-scale temporal dynamics by decoupling endogenous and exogenous signals and coupling them through structured cross-attention fusion. Specifically, a block-based encoder with a learnable global token captures both local fluctuations and global trends, while a multi-scale decomposition module extracts trend and seasonal components from key meteorological inputs such as wind speed and temperature. Furthermore, a dual cross-attention mechanism enables hierarchical interactions between endogenous and exogenous representations, enhancing representation capability. Extensive experiments on real-world datasets demonstrate that MS-ECFormer consistently outperforms state-of-the-art baselines in both predictive accuracy and stability. Our code is available at: https://github.com/liuliu20000611-boop/MS-ECFormer.**

***Index Terms*—Wind power forecasting, Multi-scale time series modeling, Endogenous–exogenous modeling, Transformer**

I. Introduction

In the ongoing global transition toward low-carbon and sustainable power systems, wind energy has become a key component of the renewable energy portfolio due to its cleanliness and abundance [1]. However, wind power output is highly sensitive to meteorological conditions. Wind speed fluctuations and wind direction changes lead to strong intermittency and volatility [2], which complicate real-time grid scheduling and may threaten voltage and frequency stability. Therefore, accurate wind power forecasting is essential for dispatch centers to anticipate generation trends, optimize coordinated scheduling among multiple energy sources, and mitigate power fluctuations for secure and economical system operation [3].

Over the past decades, numerous forecasting approaches have been developed, including physics-based models, statistical methods, and deep learning techniques. Despite continuous progress, achieving accurate and robust forecasts remains challenging due to two key issues: (i) wind power series exhibit strong non-stationarity with multi-scale temporal patterns, where short-term fluctuations coexist with long-term trends; and (ii) forecasting depends on the coupled dynamics between endogenous signals (historical power output) and exogenous meteorological variables such as wind speed and temperature [4]. These challenges motivate models capable of jointly modeling multi-scale temporal dynamics and endogenous–exogenous interactions.

Traditional physics-based methods combine numerical weather prediction (NWP) outputs with terrain information and turbine parameters [5]. Although physically interpretable, they are computationally expensive and sensitive to initial conditions, and NWP errors may accumulate and degrade forecasting accuracy, especially for longer horizons [6]. Statistical methods such as ARIMA, SARIMA, and Kalman filtering are efficient and suitable for short-term forecasting [7], but their linearity or stationarity assumptions limit their ability to handle nonlinear and rapidly changing wind conditions [8].

Recently, deep learning, especially Transformer-based architectures, has shown strong capability in modeling long-range dependencies through self-attention mechanisms [9]–[11]. However, directly applying Transformers to wind power forecasting remains challenging. Many models struggle to capture both local volatility and global trends simultaneously, leading to insufficient modeling of multi-scale non-stationary dynamics. Moreover, most methods treat endogenous power data and exogenous meteorological variables as homogeneous inputs, ignoring the asymmetric driving relationship in which meteorological factors act as external forcing variables, which may lead to feature interference and representation misalignment under highly variable meteorological conditions [13].

To address these issues, we propose MS-ECFormer, a Multi-

*Corresponding author: Zheng Ye (email: yezh0716@gmail.com)

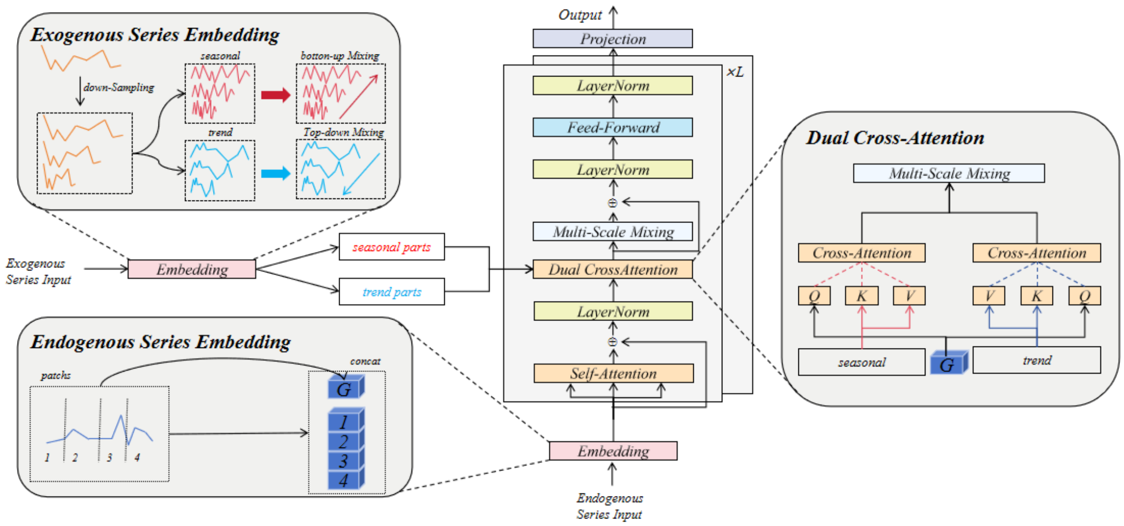

Fig. 1. Architectural schematic of the proposed MS-ECFormer.

Scale Endogenous–Exogenous Coupling Transformer for wind power forecasting. The model explicitly separates endogenous and exogenous information and models their interactions within a unified Transformer framework, enabling improved representation of non-stationary dynamics and heterogeneous dependencies in wind power series.

The main contributions are summarized as follows:

1) We propose MS-ECFormer, a Transformer-based forecasting model that decouples endogenous and exogenous variables and models their interactions through structured coupling for wind power forecasting.
2) We develop a multi-scale modeling strategy that captures local fluctuations and long-term trends by combining patch-based endogenous encoding with multi-scale decomposition of exogenous meteorological variables.
3) We design a dual cross-attention module to enable hierarchical interactions between endogenous representations and multi-scale exogenous components, improving robustness and forecasting accuracy under non-stationary conditions.
4) Extensive experiments on real-world wind power datasets and time-series benchmarks show that MS-ECFormer consistently outperforms state-of-the-art methods in both accuracy and stability.

II. PROBLEM FORMULATION

In wind power forecasting, both endogenous and exogenous variables are taken into account. The endogenous variable represents the forecasting target (*e.g.*, power output), while exogenous variables provide complementary information such as wind speed and temperature. Formally, let the historical endogenous sequence be $\mathbf{X}_{1:T} = [x_1, \ldots, x_T]^\top \in \mathbb{R}^{T\times 1}$, where x_i denotes the value of the endogenous variable at time step i. The corresponding historical observations of K exogenous variable matrix are denoted as $\mathbf{Z}_{1:T} = [\mathbf{z}_1, \ldots, \mathbf{z}_T]^\top \in \mathbb{R}^{T\times K}, \mathbf{z}_t \in \mathbb{R}^{1\times K}$. Given these inputs, the forecasting objective is to predict the next S steps of the endogenous variable, which can be formulated as:

$$\hat{\mathbf{X}}_{T+1:T+S} = F(\mathbf{X}_{1:T}, \mathbf{Z}_{1:T}), \tag{1}$$

where $F(\cdot)$ denotes the proposed model.

III. METHODOLOGY

The overall framework of MS-ECFormer adopts a pure encoder-based Transformer architecture, as illustrated in Fig. 1. To effectively capture heterogeneous temporal dependencies, the model employs differentiated embedding strategies to separately process endogenous and exogenous variables, thereby explicitly disentangling internal and external representations. Temporal dynamics are modeled through self-attention, while a dual cross-attention mechanism captures hierarchical interactions between endogenous and exogenous factors. This design ensures the representation of multi-scale dynamic features, ultimately enabling robust and accurate wind power forecasting.

A. Endogenous Embedding

Endogenous variables, as the forecasting target, require modeling both short-term fluctuations and long-term dependencies. However, conventional Transformer-based models face two major limitations when processing long time series: (1) self-attention is computationally expensive for long sequences and can become inefficient or noisy when capturing long-range dependencies under strong non-stationarity, and (2) the quadratic computational complexity with respect to sequence length results in substantial memory overhead, slower training and inference, and limited scalability in practical large-scale forecasting scenarios.

To address these issues, we adopt a patch-based representation strategy, where the original time series $\mathbf{X} \in \mathbb{R}^{T\times 1}$ is divided into N consecutive patches:

$$\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N\} = \mathrm{Patch}(\mathbf{X}), \tag{2}$$

where $N = \left\lfloor \frac{T}{p} \right\rfloor$, p is the patch length, and $\mathbf{P}_i \in \mathbb{R}^{p\times 1}$ denotes the i-th patch. We pad the sequence to the nearest multiple of p and reshape it into N patches. Each patch is embedded as:

$$\mathbf{H}_{\mathrm{en}} = \mathrm{En\_embedding}(\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N), \tag{3}$$

where $\mathbf{H}_{\mathrm{en}} \in \mathbb{R}^{N\times d_h}$ represents the patch-level embeddings of endogenous variables. In practice, the embedding uses a linear projection followed by a dropout layer to prevent overfitting and improve generalization. By encoding fixed-length local windows, this approach captures short-term variations while mitigating the computational burden associated with quadratic self-attention.

To preserve global contextual information and maintain consistency with exogenous variables encoded at different granularities, we introduce a learnable global token [14]: $\mathbf{G} \in \mathbb{R}^{1\times d_h}$, which serves as a global embedding summarizing the entire sequence. The final endogenous representation is obtained by concatenating the global token with the patch embeddings:

$$\mathbf{H}'_{\mathrm{en}} = \mathrm{concat}(\mathbf{G}, \mathbf{H}_{\mathrm{en}}), \tag{4}$$

where $\mathrm{concat}(\cdot,\cdot)$ denotes concatenation along the temporal dimension. This design allows the model to jointly capture both short-term fluctuations and long-term trends while aligning the endogenous representation with the multi-scale exogenous embeddings.

B. Exogenous Embedding

Exogenous variables provide complementary information to support the forecasting of endogenous targets. However, their effectiveness is often limited by the non-stationary nature of time series, which may contain trend, seasonal, and irregular components across multiple temporal scales. Conventional models struggle to extract, disentangle, and align these heterogeneous multi-scale patterns with the evolving endogenous dynamics, thereby constraining the effective utilization, interpretability, and robustness of exogenous information in complex forecasting scenarios.

To address this, we adopt a multi-scale decomposition strategy. Given a historical exogenous sequence $\mathbf{Z} \in \mathbb{R}^{T\times K}$, we downsample it into multiple temporal resolutions via average pooling:

$$\mathbf{Z}' = \{\mathbf{Z}_m\}_{m=0}^{M}, \tag{5}$$

where $\mathbf{Z}_m \in \mathbb{R}^{\frac{T}{2^m}\times K}$, $m \in \{0, 1, \ldots, M\}$. Here, $\mathbf{Z}_0$ preserves fine-grained temporal details, while $\mathbf{Z}_M$ captures coarse, long-term dynamics. Each resolution is projected into a higher-dimensional space using a linear layer followed by dropout to prevent overfitting:

$$\mathbf{h}_{\mathrm{ex},m} = \mathrm{Dropout}(\mathrm{Linear}(\mathbf{Z}_m)). \tag{6}$$

Since each scale exhibits both trend and seasonal patterns, we further decompose the embeddings using moving averages [15]:

$$\mathbf{S}_m, \mathbf{T}_m = \mathrm{SeriesDecomp}(\mathbf{h}_{\mathrm{ex},m}), \tag{7}$$

where $\mathbf{S}_m$ and $\mathbf{T}_m$ denote the seasonal and trend components, respectively. Multi-scale interactions are then realized via bottom-up and top-down mixing strategies with downsampling and upsampling [16]:

$$m = 0 \ldots M-1, \mathbf{S}_m \leftarrow \mathbf{S}_m + \mathrm{Down}(\mathbf{S}_{m-1}), \tag{8}$$

$$m = M-1 \ldots 0, \mathbf{T}_m \leftarrow \mathbf{T}_m + \mathrm{Up}(\mathbf{T}_{m+1}). \tag{9}$$

This design enables the model to capture both high-frequency fluctuations and long-term trends of exogenous variables. By combining multi-scale decomposition with bottom-up and top-down mixing, the exogenous embeddings are effectively aligned with endogenous dynamics, facilitating richer feature interactions in subsequent dual cross-attention modules.

C. Dual Cross-Attention

To enable effective integration between endogenous and exogenous variables, MS-ECFormer employs a stepwise dual cross-attention mechanism. After the endogenous sequence is refined through self-attention, a global representation $\mathbf{G}^{\,l}$ is obtained, which serves as the query for interactions with multi-scale exogenous features.

In the cross-attention stage [17], the global endogenous query interacts separately with the seasonal and trend components of the exogenous embeddings at each scale m:

$$\begin{aligned}\left[\mathbf{S}^{\,l}_{\mathrm{mix},m}, \mathbf{T}^{\,l}_{\mathrm{mix},m}\right] = \mathrm{Attention}\Big(&Q = \mathbf{G}^{\,l},\\ &K = [\mathbf{S}_m, \mathbf{T}_m],\\ &V = [\mathbf{S}_m, \mathbf{T}_m]\Big),\end{aligned} \tag{10}$$

where $\mathbf{S}^{\,l}_{\mathrm{mix},m}$ and $\mathbf{T}^{\,l}_{\mathrm{mix},m}$ represent the attention-enhanced seasonal and trend signals, respectively. This design allows the model to selectively incorporate external information at multiple temporal scales while preserving the endogenous context.

The attended seasonal and trend components are then fused with the global endogenous marker to produce the updated endogenous representation:

$$\tilde{\mathbf{G}}^{\mathbf{l}} = \mathrm{LayerNorm}\Big(\mathbf{G}^{\,l} + \sum_{m=0}^{M} \mathbf{S}^{\,l}_{\mathrm{mix},m} + \sum_{m=0}^{M} \mathbf{T}^{\,l}_{\mathrm{mix},m}\Big), \tag{11}$$

where the summation over scales ensures that information from all temporal resolutions is aggregated.

By harmonizing endogenous self-dependencies with exogenous guidance across multiple scales, the dual cross-attention mechanism enhances the model's ability to capture complex interactions and non-stationary patterns, thereby supporting robust and accurate forecasting under varying meteorological conditions.

TABLE I
COMPARISON OF WIND POWER GENERATION PREDICTION PERFORMANCE ON SDWPF AND WSTD2 DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN RED, AND THE SECOND-BEST RESULTS ARE UNDERLINED. A LOWER MSE, MAE, OR RMSE INDICATES BETTER PERFORMANCE.

Models		MS-ECFormer			TimeXer			iTransformer			TimesNet			TimeMixer		
Metric		MSE	MAE	RMSE	MSE	MAE	RMSE	MSE	MAE	RMSE	MSE	MAE	RMSE	MSE	MAE	RMSE
SDWPF	72	0.365	0.214	0.605	0.371	0.214	0.609	0.426	0.219	0.653	0.368	0.219	0.607	0.380	0.209	0.616
	144	0.513	0.255	0.716	0.528	0.257	0.726	0.585	0.256	0.765	0.585	0.256	0.765	0.526	0.249	0.725
	216	0.608	0.279	0.780	0.617	0.280	0.785	0.705	0.284	0.840	0.610	0.279	0.785	0.631	0.275	0.794
	288	0.646	0.294	0.804	0.661	0.295	0.813	0.751	0.297	0.867	0.655	0.291	0.810	0.664	0.289	0.815
	AVG	0.533	0.261	0.726	0.544	0.262	0.733	0.617	0.264	0.781	0.545	0.261	0.735	0.550	0.256	0.738
WSTD2	12	1.736	1.043	1.198	1.860	1.061	1.237	2.083	1.108	1.370	1.973	1.100	1.336	2.105	1.121	1.194
	24	3.943	1.577	1.784	3.899	1.581	1.778	4.075	1.598	1.905	4.910	1.691	2.073	4.520	1.649	1.741
	48	8.706	2.372	2.633	8.769	2.385	2.646	10.093	2.460	2.940	10.983	2.553	3.086	11.313	2.359	2.495
	96	18.797	3.471	3.832	18.784	3.480	3.837	19.977	3.601	3.941	18.601	3.401	3.986	21.438	3.477	3.663
	AVG	8.296	2.116	2.362	8.328	2.127	2.375	9.057	2.192	2.539	9.117	2.186	2.620	9.844	2.152	2.273

TABLE II
PERFORMANCE COMPARISON OF GENERAL SEQUENCE PREDICTION MODELS BASED ON THE ETT DATASET. ALL EXPERIMENTS EMPLOYED A 96-STEP HISTORICAL DATA LENGTH WITH A PREDICTION LENGTH OF 96. OPTIMAL RESULTS ARE HIGHLIGHTED IN RED. LOWER MEAN SQUARED ERROR (MSE) OR MEAN ABSOLUTE ERROR (MAE) VALUES INDICATE SUPERIOR PERFORMANCE.

Models	MS-ECFormer		TimeXer		iTransformer		TimesNet		TimeMixer		RLinear		PatchTST	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.374	0.399	0.382	0.403	0.386	0.405	0.384	0.402	0.375	0.400	0.386	0.395	0.414	0.419
ETTh2	0.283	0.335	0.296	0.338	0.297	0.349	0.340	0.374	0.289	0.341	0.288	0.338	0.302	0.348
ETTm1	0.309	0.355	0.318	0.356	0.334	0.368	0.338	0.375	0.320	0.357	0.355	0.376	0.329	0.367
ETTm2	0.173	0.256	0.171	0.256	0.180	0.264	0.187	0.267	0.175	0.258	0.182	0.265	0.175	0.259

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: In wind power forecasting tasks, we use two professional datasets: the SDWPF dataset from the KDD Cup [18] and the Wind Spatio-Temporal Dataset 2 (WSTD2) [19]. To further evaluate the model's generalization in capturing temporal dependencies, we also conduct experiments on four widely used ETT benchmarks (ETTh1/2 and ETTm1/2). To prevent data leakage and simulate real-world scenarios, all datasets are split into training, validation, and testing sets in a 7:1:2 ratio based on chronological order.

2) Baselines: We adopt six well-established state-of-the-art time series forecasting models as baselines: TimeXer [14], iTransformer [21], TimesNet [22], TimeMixer [16], RLinear [23], and PatchTST [12].

3) Implementation Details: All implementations are developed in PyTorch [20] and executed on a single NVIDIA RTX A6000 GPU. The model is trained using the L_2 loss, with hyperparameters adaptively configured according to dataset characteristics and sequence length to balance accuracy and computational efficiency. We use Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) as evaluation metrics.

B. Experimental Results

1) Wind Power Forecasting: Table I summarizes the quantitative forecasting performance of all compared methods on two real-world wind power datasets, SDWPF and WSTD2, under multiple prediction horizons. Overall, MS-ECFormer achieves the best average performance on both datasets, particularly in terms of MSE and MAE, demonstrating its effectiveness in modeling complex wind power dynamics.

On SDWPF, MS-ECFormer obtains the lowest errors across all horizons (72/144/216/288) for MSE and RMSE, achieving the best average MSE (0.533) and RMSE (0.726). Compared with the strongest baseline, TimeXer, the average MSE is reduced from 0.544 to 0.533. Moreover, the advantage remains stable as the forecasting horizon increases, indicating that the endogenous–exogenous coupling design helps alleviate error accumulation in long-term prediction.

On WSTD2, MS-ECFormer achieves the best average MSE (8.296) and MAE (2.116), showing robust performance under highly volatile wind power conditions. Although some baselines achieve the best results on specific horizons or metrics, MS-ECFormer maintains competitive performance across all horizons and provides the best overall accuracy in terms of MSE and MAE. These results demonstrate that combining multi-scale exogenous representations with structured endogenous–exogenous coupling improves forecasting reliability across different horizons.

2) General-purpose sequence prediction: To further evaluate the generalization ability of MS-ECFormer beyond wind power forecasting, we test the model on four widely used ETT benchmarks (ETTh1, ETTh2, ETTm1, and ETTm2) with a fixed setting of 96-step input and 96-step prediction horizon, as shown in Table II. Across these benchmarks, MS-ECFormer demonstrates strong predictive performance. It achieves the best MSE on ETTh1, ETTh2, and ETTm1, and remains competitive on ETTm2, where it obtains the second-best MSE

TABLE III
ABLATION STUDY OF MS-ECFORMER ON WSTD2 DATASETS.

Methods	MSE	MAE	RMSE
MS-ECFormer	1.736	1.043	1.198
wo/ exogenous variable_trend	1.826	1.057	1.227
wo/ exogenous variable_season	1.865	1.075	1.246
w/o Dual Cross-Attention	1.852	1.067	1.240
w/o Endogenous G	1.917	1.111	1.250
w/ endogenous variable_season	1.788	1.071	1.224
w/ endogenous variable_trend	1.817	1.072	1.231
w/ endogenous variable_trend + w/ endogenous variable_season	1.858	1.070	1.241

(0.173) with a tied MAE (0.256).Compared with TimeXer, MS-ECFormer reduces MSE from 0.382 to 0.374 on ETTh1 and from 0.296 to 0.283 on ETTh2, and also achieves the best result on ETTm1 (0.309 vs. 0.318). These results indicate that the proposed multi-scale coupling design is effective not only for wind power forecasting but also for general multivariate time-series forecasting tasks.

C. Ablation Study

To verify the effectiveness of the core components in MS-ECFormer, we conduct ablation experiments on the WSTD2 dataset. As shown in Table III, the Exogenous Disentanglement Module plays a key role in improving predictive performance. When the trend or seasonal decomposition of exogenous variables is removed, the MSE increases from 1.736 to 1.826 and 1.865, respectively, indicating that multi-scale feature extraction from meteorological variables helps mitigate non-stationarity. Furthermore, replacing the Dual Cross-Attention with a single cross-attention or removing the Endogenous Global Token increases the MSE to 1.852 and 1.917, showing that global context–guided hierarchical endogenous–exogenous interactions are important for capturing complex wind dynamics.

We also investigate applying a similar decomposition strategy to endogenous variables. The results show that decomposing the endogenous power series does not improve performance and may even reduce accuracy, with MSE increasing to 1.788–1.858. A possible reason is that endogenous variables already contain the primary dynamics of the target series, and excessive disentanglement may disrupt their integrity and introduce redundant representations. Therefore, MS-ECFormer only disentangles exogenous inputs, which enhances the complementary information from meteorological variables while preserving the coherence of endogenous features.

D. Model Analysis

1) Varying Input Sequence Length: We evaluate the impact of different input sequence lengths ($seq_len \in 36, 72, 108, 144$) under a fixed prediction horizon on the SDWPF dataset. As shown in Fig. 2, MS-ECFormer maintains a competitive advantage, achieving superior accuracy in most settings and stable performance as the input length increases. In contrast,

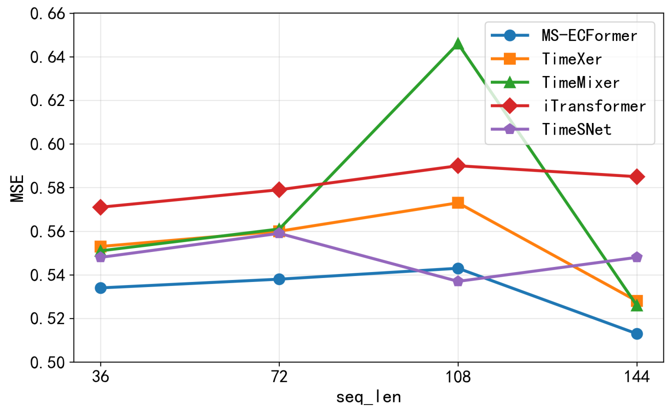

Fig. 2. Comparison of prediction errors (MSE) across different input lengths (36, 72, 108, 144) for each comparison model when the prediction length is fixed at 144.

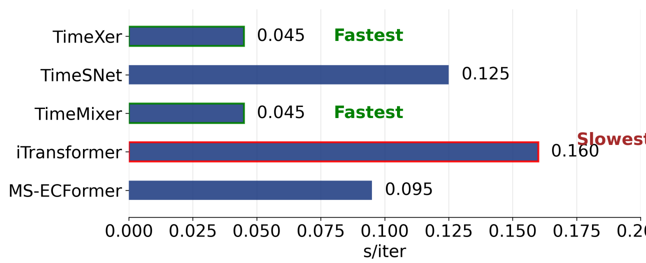

Fig. 3. Average time per iteration (seconds per iteration) measured under identical hardware configurations.In the SDWPF dataset, the input length for all baseline models is set to 144, with a prediction length of 288.

baseline models such as TimeMixer and iTransformer show noticeable performance fluctuations, especially with longer input sequences, indicating higher sensitivity to redundant or noisy historical information. This robustness is mainly attributed to the patch-based endogenous embedding and the multi-scale coupling mechanism, which help mitigate attention diffusion and selectively integrate informative patterns. Overall, these results validate the stability and scalability of MS-ECFormer across different look-back windows.

2) Efficiency Comparison: To evaluate the practical efficiency of our model, we measure the average time per iteration of MS-ECFormer and representative baselines under identical hardware settings. As shown in Fig. 3, MS-ECFormer achieves an iteration speed of 0.095 s/iter, demonstrating a good balance between predictive performance and computational efficiency. Although TimeXer and TimeMixer are the fastest (0.045 s/iter), MS-ECFormer is still more efficient than iTransformer (0.160 s/iter) and TimesNet (0.125 s/iter). This efficiency is mainly attributed to the patch-based input representation, which reduces the effective sequence length and alleviates the computational cost of attention operations under long prediction horizons. Overall, MS-ECFormer improves forecasting accuracy through multi-scale interactions while maintaining reasonable computational overhead, making it suitable for real-time wind power forecasting applications.

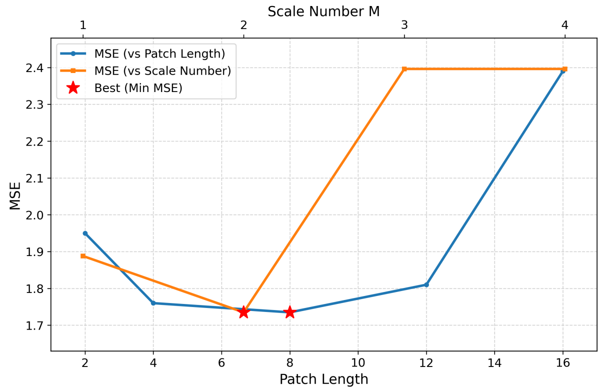

Fig. 4. Sensitivity analysis of the patch length p and downsampling layers M on the WSTD2 dataset (Input=48, Predict=12).

3) Hyper-parameter Analysis: We evaluate the sensitivity of MS-ECFormer to two core hyper-parameters: the patch length p and the number of downsampling layers M, under a forecasting setting of a 48-step historical look-back window and a 12-step prediction horizon on the WSTD2 dataset.

As illustrated in Fig. 4, the MSE exhibits a U-shaped trend as the patch length p varies in $4, 6, 8, 12, 16$, with the best performance at $p = 8$, which partitions the 48-step input into six meaningful patches. When p is too small (e.g., 4), the model becomes sensitive to high-frequency noise, while larger values (e.g., 16) lead to increased MSE due to over-smoothing of critical temporal variations and failure to capture non-stationary transitions. This indicates that an appropriate patch size is essential to balance local detail preservation and global pattern abstraction.

We further analyze the impact of the number of downsampling layers $M \in 1, 2, 3, 4$. The results show that $M = 2$ achieves the lowest MSE. A shallow structure ($M = 1$) cannot fully model multi-scale endogenous–exogenous interactions, whereas larger values ($M = 3$ or 4) degrade performance due to information dilution caused by excessive downsampling under the short input length (48 steps). This suggests that moderate decomposition depth is necessary to effectively capture multi-scale dependencies without losing fine-grained information. Therefore, $p = 8$ and $M = 2$ are selected as the optimal configuration for robust feature extraction and coupling.

V. Conclusion

In this paper, we propose MS-ECFormer, a multi-scale endogenous–exogenous coupling Transformer for wind power forecasting. By disentangling endogenous and exogenous variables, the model captures local fluctuations and global trends while extracting multi-scale seasonal and trend components from meteorological inputs. A dual cross-attention mechanism further enables hierarchical interactions between endogenous and exogenous variables, improving adaptability under non-stationary conditions. Future work will focus on improving forecasting performance under extreme meteorological events using more advanced modeling techniques.

References

[1] Wiser R, Lantz E, Mai T, et al. Wind vision: A new era for wind power in the United States[J]. The Electricity Journal, 2015, 28(9): 120-132.
[2] Hong T, Pinson P, Fan S. Global energy forecasting competition 2012[J]. International Journal of Forecasting, 2014, 30(2): 357-363.
[3] Chen N, Qian Z, Nabney I T, et al. Wind power forecasts using Gaussian processes and numerical weather prediction[J]. IEEE Transactions on Power Systems, 2013, 29(2): 656-665.
[4] Wang H, Li G, Wang G, et al. Deep learning based ensemble approach for probabilistic wind power forecasting[J]. Applied energy, 2017, 188: 56-70.
[5] Anaya-Lara O, Jenkins N, Ekanayake J B, et al. Wind energy generation: modelling and control[M]. John Wiley & Sons, 2011.
[6] Costa A, Crespo A, Navarro J, et al. A review on the young history of the wind power short-term prediction[J]. Renewable and Sustainable Energy Reviews, 2008, 12(6): 1725-1744.
[7] Zou F, Fu W, Fang P, et al. A hybrid model based on multi-stage principal component extraction, GRU network and KELM for multi-step short-term wind speed forecasting[J]. IEEE Access, 2020, 8: 222931-222943.
[8] Sideratos G, Hatziargyriou N D. An advanced statistical method for wind power forecasting[J]. IEEE Transactions on power systems, 2007, 22(1): 258-265.
[9] Chen G, Tang B, Zeng X, et al. Short-term wind speed forecasting based on long short-term memory and improved BP neural network[J]. International Journal of Electrical Power & Energy Systems, 2022, 134: 107365.
[10] Zhang Y, Yan J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting[C]//The eleventh international conference on learning representations. 2023.
[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[12] Huang X, Tang J, Shen Y. Long time series of ocean wave prediction based on PatchTST model[J]. Ocean Engineering, 2024, 301: 117572.
[13] Wu Z, Luo G, Yang Z, et al. A comprehensive review on deep learning approaches in wind forecasting applications[J]. CAAI Transactions on Intelligence Technology, 2022, 7(2): 129-143.
[14] Wang Y, Wu H, Dong J, et al. Timexer: Empowering transformers for time series forecasting with exogenous variables[J]. Advances in Neural Information Processing Systems, 2024, 37: 469-498.
[15] Cleveland R B, Cleveland W S, McRae J E, et al. STL: A seasonal-trend decomposition[J]. J. off. Stat, 1990, 6(1): 3-73.
[16] Wang S, Wu H, Shi X, et al. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting[C]//The Twelfth International Conference on Learning Representations.
[17] Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for computational linguistics. Meeting. 2019, 2019: 6558.
[18] Zhou J, Lu X, Xiao Y, et al. Sdwpf: A dataset for spatial dynamic wind power forecasting challenge at kdd cup 2022[J]. arXiv preprint arXiv:2208.04360, 2022.
[19] Y. Ding, Data Science for Wind Energy. Boca Raton, FL: Chapman & Hall/CRC Press, 2019.
[20] Imambi S, Prakash K B, Kanagachidambaresan G R. PyTorch[M]//Programming with TensorFlow: solution for edge computing applications. Cham: Springer International Publishing, 2021: 87-104.
[21] Liu Y, Hu T, Zhang H, et al. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting[C]//The Twelfth International Conference on Learning Representations.
[22] Wu H, Hu T, Liu Y, et al. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis[C]//The Eleventh International Conference on Learning Representations.
[23] Li Z, Qi S, Li Y, et al. Revisiting long-term time series forecasting: An investigation on linear mapping[J]. arXiv preprint arXiv:2305.10721, 2023.