Efficiently Enhancing Long-term Series Forecasting via Adaptive Lookback with Wavelets

Suxin Tong, Jingling Yuan*

Hubei Key Laboratory of Transportation Internet of Things, School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China
suxin_tong@whut.edu.cn, yjl@whut.edu.cn

Abstract

Long-term series forecasting leverages historical observations to predict extended future sequences and plays a crucial role across various domains. However, conventional models relying on fixed-length lookback windows struggle with the inherent dynamic dependencies and multi-scale characteristics of time series data. These fixed windows either introduce noise through excessive length or omit critical patterns when too short, while optimal window sizes vary significantly with tasks and external conditions. To address this, we propose the Adaptive Lookback Window (ALW) framework, a wavelet transform-driven approach for multi-scale adaptive lookback window selection. ALW decomposes time series into distinct frequency components through wavelet transforms, quantifies the contribution of each historical time step via scale-specific attention mechanisms, and dynamically determines optimal window lengths through backward information accumulation and soft truncation techniques. Finally, refined input features are generated for downstream prediction models via a weighted reconstruction process. Extensive evaluations across multiple public benchmarks demonstrate that ALW, as an efficient plug-and-play technique, not only reduces the MSE of backbone models by an average of 3.2% but also alleviates hyperparameter tuning requirements and enables input feature dimensionality reduction, which curtails subsequent model computational costs.

Code — https://github.com/SuxinTong/ALW

Introduction

Long-term series forecasting is crucial for data-driven decision-making in fields like power load forecasting (Uremovic et al. 2023; Zhong et al. 2024), traffic planning (Jiang et al. 2023; Li et al. 2024), and meteorological prediction (Huang et al. 2023; Luo et al. 2023). Its accuracy directly enables resource optimization and risk mitigation, representing significant economic and social value. Recently, deep learning models such as Convolutional Neural Networks(Liu et al. 2022a; Wu et al. 2023; Wang et al. 2023), Multi-Layer Perceptrons(Zeng et al. 2023; Wang et al. 2024; Lin et al. 2024) and Transformers (Wu et al. 2021; Zhou et al. 2022; Nie et al. 2023; Liu et al. 2024; Qiu et al. 2025; Lin et al. 2025), have achieved remarkable progress in time series forecasting. However, these models predominantly rely on fixed-length lookback windows determined either a priori or through hyperparameter search.

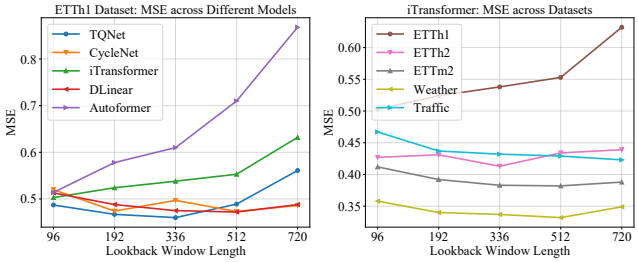

Figure 1: The optimal lookback window length typically varies across different models and datasets. Forecasting performance with fixed prediction length $F = 720$.

As illustrated in Figure 1, optimal lookback window length varies significantly across different models (Wu et al. 2021; Zeng et al. 2023; Liu et al. 2024; Lin et al. 2024, 2025) and datasets, revealing fundamental limitations of fixed-length strategies. Firstly, fixed windows neglect the dynamic nature of temporal dependencies in real-world time series, where information relevance fluctuates over time. Stable periods may require only short-term data, while post-event recovery periods necessitate longer windows to capture causal relationships. Consequently, excessive window length introduces noise or outdated information, while insufficient length loses long-term trends or critical patterns. Secondly, time series typically contain patterns across various temporal scales, such as intra-day fluctuations, weekly seasonality, and annual trends in power load data. A single fixed window cannot simultaneously optimize the capture of all scale patterns. Finally, fixed window approaches require computationally expensive hyperparameter searches to determine globally optimal lengths while lacking sample-specific adaptive capabilities. These limitations severely restrict model generalization and practical application.

To our knowledge, a universal, end-to-end, data-driven adaptive lookback window framework remains absent. Given the aforementioned limitations, an ideal historical lookback mechanism should be universal across domains

*Corresponding author.

and model architectures, dynamically determining an optimal window range through an end-to-end differentiable process that is jointly optimized with the forecasting backbone, thereby maximizing relevant information while minimizing noise and outdated data. Attention-based models generate signals measuring associations between different time steps. By distinguishing these signals as positive or negative and accumulating them from the most recent time step backward, the distribution of the cumulative curve informs the historical point beyond which additional context yields diminishing predictive returns, thus defining a soft boundary of effective history. The challenge lies in transforming this information importance accumulation process into an end-to-end learnable, differentiable window selection mechanism that preserves the continuous structure of time series rather than selecting discrete points. Additionally, considering the multi-scale characteristics of time series, this adaptive mechanism should operate independently at different temporal granularities to precisely capture critical historical information required for patterns at various scales.

Motivated by these insights, we propose the **A**daptive **L**ookback **W**indow (ALW) framework, driven by wavelet transforms for multi-scale analysis. ALW comprises three core modules: Multi-scale Decomposition Module, Adaptive Lookback Window Learning Module, and Weighted Reconstruction Module. The Multi-scale Decomposition Module uses wavelet transforms to decompose the input time series into multiple components representing different frequency constituents, each carrying pattern information at specific temporal scales. The Adaptive Lookback Window Learning Module evaluates the importance of each time step through scale-specific attention mechanisms for each frequency component and dynamically determines the most relevant historical lookback range through backward information accumulation and soft truncation techniques. Finally, the Weighted Reconstruction Module transforms and fuses the features processed by adaptive windows at various scales and reconstructs them into input features suitable for downstream prediction models. Through these mechanisms, ALW achieves adaptive selection and utilization of the most relevant historical information at multiple scales, overcoming the limitations of conventional fixed window methods.

The main contributions of this paper include:

- We propose a novel wavelet transform-driven multi-scale adaptive lookback window framework (ALW) that dynamically learns optimal lookback lengths for each instance and scale, effectively addressing the dynamic and multi-scale nature of time series.
- We design a differentiable window selection mechanism based on cumulative information contribution and soft truncation, enabling end-to-end optimization of lookback window length while providing interpretability for the window selection process.
- Experiments on multiple public benchmarks demonstrate that ALW reduces backbone models' MSE by an average of 3.2%, alleviates hyperparameter tuning requirements, and enables input feature dimensionality reduction to lower computational costs.

Related Work

Time Series Forecasting

Time series forecasting has progressed from traditional statistical methods (Farlie 1964; Watson 1994; Makridakis and Hibon 1997) to deep learning techniques such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), Transformers, and multilayer perceptrons (MLPs). RNNs (Hochreiter and Schmidhuber 1997; Cao et al. 2018; Baytas et al. 2017) capture sequential dependencies via recurrent states, while CNNs (van den Oord et al. 2016; Bai, Kolter, and Koltun 2018) exploit temporal convolutions. However, their limited receptive fields hinder long-term forecasting. Transformer models (Vaswani et al. 2017), adapted from natural language processing, leverage attention mechanisms for time series forecasting. Variants including Autoformer (Wu et al. 2021), PatchTST (Nie et al. 2023), iTransformer (Liu et al. 2024), and TQNet (Lin et al. 2025) enhance performance through improved attention mechanisms, patching, transposition, and periodically shifted learnable queries. Recently, efficient MLP-based architectures have gained traction: DLinear (Zeng et al. 2023), FITS (Xu, Zeng, and Xu 2024), and CycleNet (Lin et al. 2024) achieve strong results with fewer parameters, while TimeMixer (Wang et al. 2024) and WaveletMixer (Zhang et al. 2025) incorporate multi-scale processing into MLP frameworks to boost performance. Despite these advances, most methods rely on a fixed-length lookback window. Consequently, we propose ALW, a lightweight, plug-and-play framework built on an MLP backbone that adaptively selects relevant history while preserving computational efficiency.

Adaptive Lookback Window for Time Series Forecasting

Selecting an appropriate lookback window is critical for accurate forecasting, as real-world time series exhibit dynamic dependencies across varying time scales. To mitigate the limitations of fixed windows, prior work has tailored window configurations to domain characteristics. In financial forecasting, John et al. (John, Binnewies, and Stantic 2025) adjust historical length by presetting different thresholds based on market volatility, and John et al. (John, Binnewies, and Stantic 2024) underscore the absence of domain-agnostic methods for optimal window selection. Similarly, Koparanov et al. (Koparanov, Georgiev, and Shterev 2020) and Tong et al. (Tong and Yuan 2025) examine how window length interacts with signal characteristics and sampling frequency. However, no prior work offers a universal, end-to-end, data-driven framework that learns adaptive lookback windows. Our ALW framework addresses this gap by dynamically selecting the most relevant historical inputs for each forecast, offering interpretable window choices and seamless integration as an efficient plugin that obviates manual tuning.

Methodology

Problem Definition

Let $\mathcal{X} \in \mathbb{R}^{E \times N}$ denote a multivariate time series with E time steps and N variables. Conventional forecasting uses

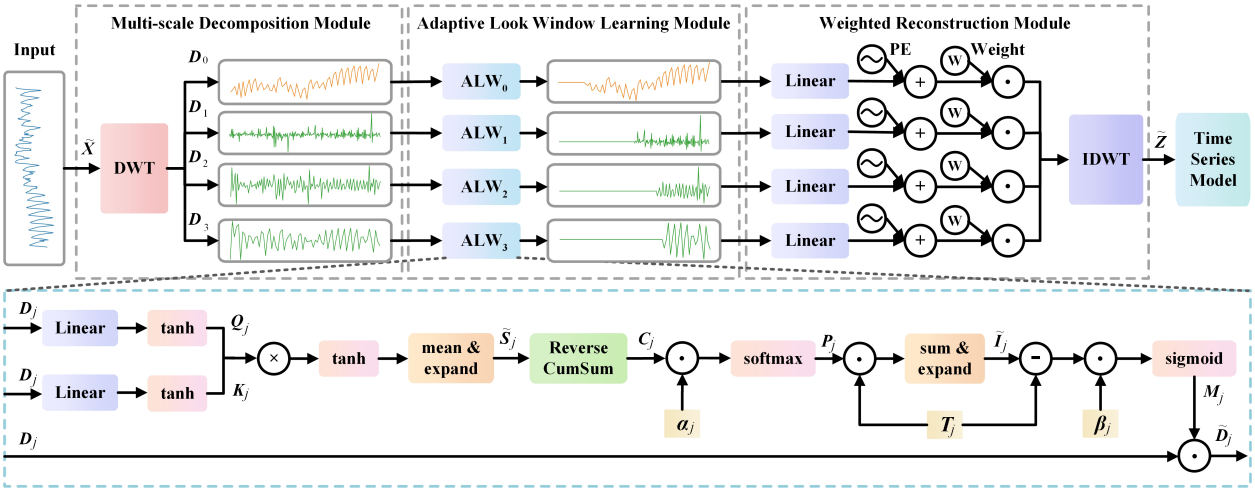

Figure 2: Architecture of the ALW framework, comprising three key modules: Multi-scale Decomposition Module, Adaptive Lookback Window Learning Module, and Weighted Reconstruction Module.

a fixed lookback window of length L to predict the next F steps. Our objective is to adaptively select a contiguous segment within the past L observations and project it to a compact representation $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{H \times N}$ (with $H < L$). A downstream model then uses $\tilde{\boldsymbol{Z}}$ to forecast $\mathcal{X}_{t+1:t+F}$. This process is formally described as $f : \mathcal{X}_{t-L+1:t} \in \mathbb{R}^{L \times N} \rightarrow \tilde{\boldsymbol{Z}} \in \mathbb{R}^{H \times N} \rightarrow \mathcal{X}_{t+1:t+F} \in \mathbb{R}^{F \times N}$.

Overall Structure

Figure 2 illustrates the overall architecture of ALW, which comprises three key components: a Multi-scale Decomposition Module utilizing the Discrete Wavelet Transform (DWT), an Adaptive Lookback Window Learning Module for determining the lookback boundaries at each scale (detailed in Figure 3), and a Weighted Reconstruction Module that integrates the multi-scale features. These components will be elaborated upon in subsequent sections.

Multi-scale Decomposition Module

Time series data inherently contains information across multiple temporal scales, including short-term fluctuations, medium-term cycles, and long-term trends. Since a fixed window struggles to capture these dynamic dependencies, we leverage DWT, which effectively localizes patterns in both time and frequency domains to capture variations at different temporal granularities. Specifically, for an input sequence $\boldsymbol{X} \in \mathbb{R}^{L \times N}$, we first perform a normalization process (Liu et al. 2022b) and transpose it to obtain $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N \times L}$. We then apply J-level wavelet decomposition to decompose $\tilde{\boldsymbol{X}}$:

$$\{\boldsymbol{D}_j\}_{j=0}^{J} = \{\boldsymbol{D}_0, \{\boldsymbol{D}_j\}_{j=1}^{J}\} = \mathrm{DWT}(\tilde{\boldsymbol{X}}, J), \quad (1)$$

where J is the decomposition level. $\boldsymbol{D}_0 \in \mathbb{R}^{N \times L_0}$ and $\{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_J\} \in \mathbb{R}^{N \times L_j}$ represent the lowest frequency (coarsest scale) and different detail levels (higher frequencies) information, respectively. The length L_j of each component depends on the original sequence length L and the decomposition level J. This decomposition separates information at different frequencies, enabling the model to learn optimal lookback windows independently across various frequency domains. Furthermore, the reduced dimensionality of components contributes to computational efficiency while providing targeted feature representations for subsequent adaptive lookback window learning.

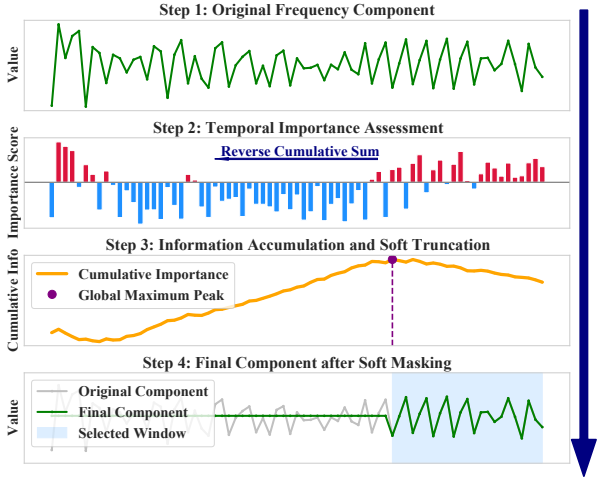

Figure 3: Overview of the Adaptive Lookback Window Learning Module.

Adaptive Lookback Window Learning Module

Conventional methods employing a fixed historical window length struggle to adapt to the dynamic dependencies inherent in time series data. An effective lookback window should capture the most relevant and informative historical context for accurate forecasting. Fixed windows or hard truncation methods lack the necessary flexibility and differentiability for end-to-end optimization through backpropagation. Fol-

lowing multi-scale decomposition, we aim to adaptively determine the optimal historical lookback window for each frequency component. To achieve this, we design a differentiable Adaptive Lookback Window Learning Module incorporating attention, information accumulation, and soft selection, as illustrated in Figure 3.

Temporal Importance Assessment For each decomposed frequency component $\boldsymbol{D}_j \in \mathbb{R}^{N\times L_j}$ ($j = 0, \ldots, J$), we project $\boldsymbol{D}_j$ through scale-specific linear layers to derive Query $\boldsymbol{Q}_j \in \mathbb{R}^{N\times L_j}$ and Key $\boldsymbol{K}_j \in \mathbb{R}^{N\times L_j}$ representations. The channel-aggregated temporal importance score $\boldsymbol{S}_j \in \mathbb{R}^{L_j}$ for scale j is computed as:

$$\begin{aligned} \boldsymbol{Q}_j &= \tanh(\text{Linear}_Q^{(j)}(\boldsymbol{D}_j)), \\ \boldsymbol{K}_j &= \tanh(\text{Linear}_K^{(j)}(\boldsymbol{D}_j)), \\ \boldsymbol{S}_j &= \text{mean}(\tanh(\frac{\boldsymbol{Q}_j^\top \boldsymbol{K}_j}{\sqrt{N}})), \end{aligned} \tag{2}$$

where $\text{Linear}_Q^{(j)}, \text{Linear}_K^{(j)}: \mathbb{R}^{N\times L_j} \to \mathbb{R}^{N\times L_j}$. $\boldsymbol{Q}_j$ is trained to summarize critical predictive features (e.g., dominant seasonal patterns), while $\boldsymbol{K}_j$ encodes features at each historical time step. Consequently, high dot-product scores indicate strong alignment between historical patterns (Key) and predictive features (Query), serving as a measure of temporal importance. The $\tanh$ function bounds scores to $(-1, 1)$, stabilizing training while enabling semantic interpretability (e.g., negative values denote inverse correlation). By averaging along the first dimension via the mean operation, we obtain a unified importance profile $\boldsymbol{S}_j$, which is then broadcast to $\tilde{\boldsymbol{S}}_j \in \mathbb{R}^{N\times L_j}$ for subsequent accumulation. This design choice enforces a shared lookback window across all channels based on the assumption that temporal dependencies are primarily driven by common underlying factors rather than channel-specific dynamics. This regularization prevents overfitting to channel-specific noise and promotes robust pattern learning, enhancing both generalization and computational efficiency.

Information Accumulation and Soft Truncation To simulate reviewing history from the current point, we perform a backward cumulative sum on the importance weights $\tilde{\boldsymbol{S}}_j$ along the time dimension, obtaining cumulative importance $\boldsymbol{C}_j \in \mathbb{R}^{N\times L_j}$. We posit that time points proximate to the peak of this cumulative contribution delineate soft boundaries for an effective historical lookback. To learn these boundaries differentiably, we employ a soft-argmax mechanism based on $\boldsymbol{C}_j$ to derive continuous soft lookback indices $\boldsymbol{I}_j \in \mathbb{R}^{N\times 1}$. This process is defined as:

$$\begin{aligned} \boldsymbol{C}_j^{(t)} &= \sum_{k=t}^{L_j-1} \tilde{\boldsymbol{S}}_j^{(k)}, \quad \boldsymbol{P}_j = \text{softmax}(\alpha_j \cdot \boldsymbol{C}_j), \\ \boldsymbol{I}_j &= \sum_{t=0}^{L_j-1} \boldsymbol{T}_j^{(t)} \odot \boldsymbol{P}_j^{(t)}, \end{aligned} \tag{3}$$

where $\boldsymbol{C}_j^{(t)}$ is the cumulative importance at time step t, and $\alpha_j > 0$ is a learnable parameter controlling the peak sharpness of the Softmax distribution. $\odot$ represents element-wise multiplication. The probability distribution $\boldsymbol{P}_j \in \mathbb{R}^{N\times L_j}$ indicates the likelihood of each historical position serving as an effective lookback cutoff point for each feature channel. By computing a weighted sum of this distribution with the time indices vector $\boldsymbol{T}_j = [\boldsymbol{0}, \boldsymbol{1}, \ldots, \boldsymbol{L}_j - 1] \in \mathbb{R}^{N\times L_j}$, we obtain the continuous, differentiable soft boundary index $\boldsymbol{I}_j$ for each feature channel. This index $\boldsymbol{I}_j$ is then broadcast to match the dimensions of $\boldsymbol{T}_j$, yielding $\tilde{\boldsymbol{I}}_j \in \mathbb{R}^{N\times L_j}$ for subsequent processing.

Soft Masking Mechanism Based on the learned soft indices $\tilde{\boldsymbol{I}}_j$, we construct adaptive soft masks $\boldsymbol{M}_j \in \mathbb{R}^{N\times L_j}$ and apply them to the original frequency components:

$$\boldsymbol{M}_j = \sigma(\beta_j \cdot (\boldsymbol{T}_j - \tilde{\boldsymbol{I}}_j)), \quad \tilde{\boldsymbol{D}}_j = \boldsymbol{D}_j \odot \boldsymbol{M}_j, \tag{4}$$

where σ is the sigmoid function, $\beta_j > 0$ is a learnable parameter controlling the mask steepness; larger values approximate hard truncation. The resulting mask $\boldsymbol{M}_j$ assigns weights close to 0 for time steps earlier than $\tilde{\boldsymbol{I}}_j$ and weights close to 1 for those after, effectively achieving smooth information truncation. Finally, the mask is applied element-wise to the original frequency component $\boldsymbol{D}_j$ to obtain the adaptively filtered component $\tilde{\boldsymbol{D}}_j$.

Weighted Reconstruction Module

The Weighted Reconstruction Module transforms and integrates adaptively processed multi-scale features into a unified representation for downstream forecasting. Specifically, we apply scale-specific linear transformations $\text{Linear}_Z^{(j)}: \mathbb{R}^{N\times L_j} \to \mathbb{R}^{N\times H_j}$ to the masked components $\{\tilde{\boldsymbol{D}}_j\}_{j=0}^J$. We then enhance the transformed features with Positional Encoding (PE) (Vaswani et al. 2017) and scale them using learnable channel importance weights. Finally, we reconstruct the unified feature representation $\tilde{\boldsymbol{Z}}$ via the Inverse Discrete Wavelet Transform (IDWT):

$$\begin{aligned} \boldsymbol{Z}_j &= (\text{Linear}_Z^{(j)}(\tilde{\boldsymbol{D}}_j) + \text{PE}(H_j)) \odot \text{softmax}(\boldsymbol{W}_Z^{(j)}), \\ \tilde{\boldsymbol{Z}} &= \text{IDWT}(\boldsymbol{Z}_0, \{\boldsymbol{Z}_j\}_{j=1}^J), \end{aligned} \tag{5}$$

where PE enhances the model's perception of temporal ordering within each scale. Although the preceding module applies uniform temporal masking across channels using a shared importance profile, channel contributions within $\tilde{\boldsymbol{D}}_j$ may vary significantly. To address this, we introduce learnable channel weights $\boldsymbol{W}_Z^{(j)} \in \mathbb{R}^{N\times H_j}$, where H_j derives from the preset mapping length H via DWT. These weights are obtained through Softmax along the channel dimension, enabling the model to adaptively emphasize reliable channels per scale and prediction time step. Finally, the reconstructed feature representation $\tilde{\boldsymbol{Z}}$ serves as the input for the downstream prediction model to forecast $\mathcal{X}_{t+1:t+F}$.

Experiments

Experimental Settings

Datasets We evaluated the proposed ALW on eight widely used public datasets from time series forecasting research (Liu et al. 2024): Weather, Exchange, ECL, Traffic, and four ETT datasets (ETTh1, ETTh2, ETTm1, ETTm2).

Model		TQNet		+ALW		CycleNet		+ALW		iTransformer		+ALW		DLinear		+ALW	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.370	0.398	**0.366**	**0.397**	0.375	0.395	**0.373**	**0.394**	0.386	0.405	**0.375**	**0.405**	0.375	0.399	**0.365**	**0.395**
	192	0.408	0.424	**0.403**	**0.422**	0.405	0.423	**0.403**	**0.422**	0.424	0.440	**0.423**	**0.437**	0.405	0.416	**0.401**	**0.416**
	336	0.433	0.441	**0.432**	**0.434**	0.426	0.436	**0.423**	**0.436**	0.459	0.459	**0.450**	**0.458**	0.439	0.443	**0.429**	**0.435**
	720	0.460	0.471	**0.446**	**0.469**	0.473	0.474	**0.448**	**0.460**	0.503	0.491	**0.486**	**0.491**	0.472	0.490	**0.465**	**0.472**
ETTh2	96	0.275	0.339	**0.274**	**0.339**	0.291	0.353	**0.271**	**0.336**	0.297	0.349	**0.285**	**0.349**	0.289	0.353	**0.277**	**0.339**
	192	0.340	0.379	**0.335**	**0.379**	0.359	0.394	**0.331**	**0.383**	0.369	0.400	**0.360**	**0.400**	0.383	0.418	**0.333**	**0.384**
	336	0.370	0.411	**0.368**	**0.410**	0.384	0.419	**0.366**	**0.416**	0.392	0.422	**0.378**	**0.422**	0.448	0.465	**0.365**	**0.415**
	720	0.394	0.434	**0.393**	**0.433**	0.417	0.446	**0.409**	**0.443**	0.413	0.442	**0.377**	**0.430**	0.605	0.551	**0.426**	**0.454**
ETTm1	96	0.291	0.343	**0.287**	**0.341**	0.297	0.351	**0.288**	**0.342**	0.316	0.366	**0.309**	**0.362**	0.299	0.343	**0.298**	**0.343**
	192	0.329	0.371	**0.329**	**0.371**	0.338	0.377	**0.329**	**0.370**	0.357	0.394	**0.350**	**0.389**	0.335	0.365	**0.332**	**0.364**
	336	0.368	0.392	**0.367**	**0.392**	0.367	0.393	**0.359**	**0.389**	0.378	0.402	**0.373**	**0.400**	0.369	0.386	**0.366**	**0.383**
	720	0.425	0.425	**0.408**	**0.423**	0.436	0.425	**0.413**	**0.416**	0.435	0.438	**0.421**	**0.430**	0.425	0.421	**0.424**	**0.416**
ETTm2	96	0.170	0.258	**0.165**	**0.255**	0.163	0.246	**0.161**	**0.245**	0.172	0.265	**0.167**	**0.261**	0.167	0.260	**0.166**	**0.259**
	192	0.226	0.298	**0.221**	**0.298**	0.228	0.293	**0.216**	**0.291**	0.249	0.314	**0.244**	**0.312**	0.224	0.303	**0.223**	**0.297**
	336	0.278	0.335	**0.276**	**0.334**	0.276	0.332	**0.268**	**0.323**	0.294	0.345	**0.294**	**0.342**	0.281	0.342	**0.278**	**0.329**
	720	0.376	0.389	**0.369**	**0.389**	0.364	0.391	**0.362**	**0.384**	0.382	0.404	**0.381**	**0.397**	0.397	0.421	**0.376**	**0.393**
Weather	96	0.153	0.205	**0.142**	**0.196**	0.148	0.200	**0.143**	**0.198**	0.163	0.213	**0.150**	**0.204**	0.176	0.237	**0.141**	**0.190**
	192	0.196	0.243	**0.185**	**0.237**	0.190	0.240	**0.188**	**0.240**	0.203	0.249	**0.194**	**0.248**	0.220	0.282	**0.185**	**0.233**
	336	0.245	0.281	**0.237**	**0.277**	0.242	0.281	**0.238**	**0.278**	0.261	0.295	**0.254**	**0.294**	0.265	0.319	**0.236**	**0.272**
	720	0.314	0.332	**0.310**	**0.329**	0.314	0.333	**0.311**	**0.332**	0.332	0.338	**0.328**	**0.337**	0.323	0.362	**0.308**	**0.325**
Exchange	96	0.084	0.202	**0.083**	**0.202**	0.084	0.202	**0.080**	**0.200**	0.086	0.206	**0.084**	**0.206**	0.082	0.207	**0.082**	**0.200**
	192	0.179	0.303	**0.167**	**0.292**	0.174	0.300	**0.173**	**0.294**	0.177	0.299	**0.169**	**0.295**	0.180	0.300	**0.178**	**0.300**
	336	0.328	0.426	**0.324**	**0.413**	0.334	0.416	**0.317**	**0.414**	0.331	0.417	**0.320**	**0.415**	0.334	0.441	**0.325**	**0.410**
	720	0.811	0.681	**0.710**	**0.640**	0.846	0.692	**0.825**	**0.681**	0.847	0.691	**0.813**	**0.680**	0.868	0.711	**0.845**	**0.689**
ECL	96	0.129	0.224	**0.129**	**0.224**	0.126	0.221	**0.126**	**0.221**	0.133	0.230	**0.130**	**0.227**	0.140	0.237	**0.132**	**0.227**
	192	0.154	0.248	**0.154**	**0.248**	0.144	0.237	**0.143**	**0.237**	0.153	0.250	**0.149**	**0.246**	0.153	0.249	**0.151**	**0.244**
	336	0.166	0.260	**0.162**	**0.260**	0.160	0.255	**0.159**	**0.254**	0.167	0.266	**0.163**	**0.260**	0.169	0.267	**0.163**	**0.256**
	720	0.201	0.294	**0.194**	**0.289**	0.199	0.291	**0.188**	**0.283**	0.195	0.289	**0.188**	**0.283**	0.203	0.301	**0.199**	**0.288**
Traffic	96	0.360	0.257	**0.359**	**0.253**	0.386	0.268	**0.384**	**0.262**	0.358	0.261	**0.350**	**0.258**	0.410	0.282	**0.397**	**0.277**
	192	0.386	0.279	**0.375**	**0.263**	0.404	0.276	**0.402**	**0.271**	0.382	0.269	**0.361**	**0.266**	0.423	0.287	**0.409**	**0.282**
	336	0.399	0.276	**0.390**	**0.272**	0.416	0.281	**0.416**	**0.277**	0.387	0.276	**0.385**	**0.276**	0.436	0.296	**0.419**	**0.287**
	720	0.447	0.315	**0.427**	**0.292**	0.445	0.300	**0.445**	**0.300**	0.429	0.299	**0.420**	**0.293**	0.466	0.315	**0.456**	**0.307**

Table 1: Comparison between backbone models with hyperparameter-searched input sequence lengths from {96, 192, 336, 512} and their ALW-enhanced versions with fixed initial input length of 512 mapped to $H = 256$. Prediction lengths $F \in \{96, 192, 336, 720\}$. Best results are shown in **bold**.

Backbone Models To evaluate ALW's performance and compatibility, we integrated it with diverse time series forecasting backbone models. These included TQNet (Lin et al. 2025) (which uses periodically shifted learnable queries in attention for global correlations, with a single-layer MHA and lightweight MLP), CycleNet-MLP (Lin et al. 2024) (a plug-and-play model), iTransformer (Liu et al. 2024) (which inverts temporal dimensionality while preserving the original Transformer architecture), and DLinear (Zeng et al. 2023) (which employs seasonal-trend decomposition).

Implementation Details All experiments were run on PyTorch 2.6.0 (Paszke et al. 2019) using two NVIDIA Tesla T4 16 GB GPUs under Linux. We optimized the L2 loss via Adam (Kingma and Ba 2015). For each backbone, we followed TimeMixer's hyperparameter search ranges (Wang et al. 2024): input lengths {96, 192, 336, 512}, learning rates from 10^{-5} to 0.05, encoder layers from 1 to 5, model dimensions from 16 to 512, training epochs from 10 to 100. For ALW-enhanced variants ("+ALW"), we fixed the input length to 512 and projected the adaptive output to 256 features before the backbone. We evaluated forecast horizons of {96, 192, 336, 720}. The DWT used the db6 wavelet with J=3. We reported MSE and MAE to capture large errors and ensure robustness to outliers, and we fixed a random seed.

Main Results

Table 1 presents a comparative performance analysis to validate ALW's effectiveness in enhancing performance and

its capacity to mitigate the typical need for hyperparameter search for input length. The comparison is between original backbones, configured with hyperparameter-searched input sequence lengths, and their ALW-enhanced versions (integrated as a plugin module), which utilized a fixed initial input length of 512. The results show that ALW consistently improved performance across eight datasets, achieving an average MSE reduction of 3.2%. Model-specific average MSE reductions were 2.3% for TQNet, 2.5% for CycleNet, 2.9% for iTransformer, and 5.1% for DLinear, highlighting ALW's broad applicability.

Notably, ALW consistently improved performance across different dataset types, with MSE reductions of 6.8% on ETTh2 and 5.7% on Weather. These datasets feature complex periodicity and trends, demanding flexible capture of multi-scale dependencies. ALW's multi-scale decomposition and adaptive lookback learning effectively identify and utilize dynamic historical information, contributing to superior performance on these benchmarks.

Furthermore, ALW maintains robust performance advantages across varying prediction horizons. This sustained performance is attributable to its capacity to adaptively adjust lookback windows, thereby incorporating sufficient relevant historical context while concurrently mitigating interference from obsolete data that might obscure critical underlying patterns. By dynamically determining relevant historical ranges, ALW provides more stable input features, particularly beneficial for long-term forecasting where dependencies are complex.

Ablation Study and Analysis

Ablation Study To ensure that performance changes can be unambiguously attributed to our modifications, we conduct an ablation study using the lightweight, plug-and-play CycleNet model as a backbone, with results on the ETTh1, ETTm1, and Weather datasets detailed in Table 2. Removing the Multi-scale Decomposition Module (w/o MSDM) impairs the model's ability to isolate distinct temporal patterns, raising MSE and MAE across all datasets. Omitting the Adaptive Lookback Window Learning Module (w/o ALWLM) degrades the mechanism for computing temporal importance and soft-truncating redundant history, similarly worsening forecasts. Excluding the Weighted Reconstruction Module (w/o WRM) forces direct inverse wavelet aggregation, which fails to transform and weight multi-scale features into a structured representation, causing a marked error increase. We further examine two variants of WRM: without positional encoding (w/o PE), which removes the sinusoidal positional embeddings that inform the reconstruction order, and without learnable weights (w/o Weights), which replaces adaptive weight parameters with uniform averaging. Both variants exhibit substantial performance declines, underscoring the necessity of PE for capturing positional context and of learnable weights for emphasizing salient scales. Collectively, these results confirm that multi-scale decomposition, adaptive window learning and weighted reconstruction synergize to deliver ALW's superior accuracy.

Dataset	ETTh1		ETTm1		Weather	
Metric	MSE	MAE	MSE	MAE	MSE	MAE
+ALW	**0.412**	**0.428**	**0.347**	**0.379**	**0.220**	**0.262**
w/o MSDM	0.420	0.433	0.359	0.386	0.240	0.280
w/o ALWLM	0.419	0.435	0.354	0.383	0.224	0.265
w/o WRM	0.448	0.453	0.366	0.391	0.224	0.266
w/o PE	0.461	0.457	0.350	0.381	0.228	0.270
w/o Weights	0.472	0.456	0.353	0.383	0.239	0.278

Table 2: Ablation study of ALW integrated into CycleNet on multiple datasets, with results averaged across all prediction lengths. Best results are shown in **bold**.

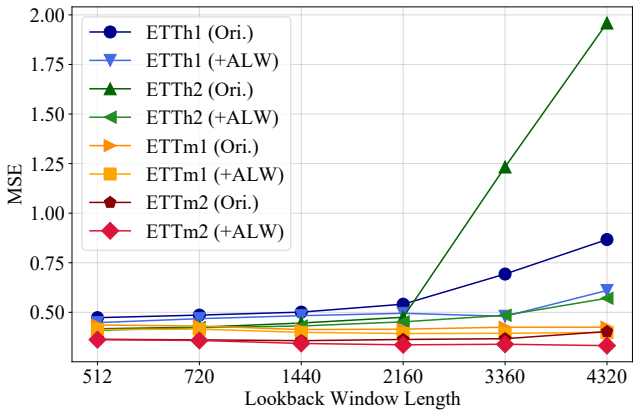

Figure 4: MSE of CycleNet (Original vs. +ALW) with increasing lookback window length on ETT datasets.

Performance with Increasing Lookback Length To investigate the impact of lookback window length, we evaluated original CycleNet (Ori.) against its ALW-enhanced version (+ALW) on four ETT datasets with prediction horizon $F = 720$, with lookback length varying from 512 to 4320 (Figure 4). For +ALW configurations, ALW consistently maps features from varying lookback windows (L) to a fixed dimension $H = 256$ before feeding to CycleNet. Results show that original CycleNet degrades with longer windows, indicating that fixed-length approaches incorporate noise or irrelevant history without adaptive filtering. While CycleNet+ALW exhibits substantially mitigated performance degradation on the hourly-sampled ETT datasets (ETTh1 and ETTh2) and sustained improvements on ETTm2, the persistent increase in MSE observed across these hourly benchmarks necessitates further analysis. This occurs due to three interrelated factors: First, extremely long windows introduce diverse temporal patterns with weak correlations to the target horizon; as L increases, the soft-argmax distribution over the cumulative importance curve becomes broader, causing ALW to include marginally relevant history despite adaptive selection. Second, the fixed projection to $H = 256$ creates a representational bottleneck that cannot fully preserve informative signals from excessively long sequences. Third, hourly transformer state data exhibits stronger non-

Wavelet Families			Decomposition Levels		
Dataset Wavelet	ETTm1 MSE	Weather MSE	Dataset Level	ETTm1 MSE	Weather MSE
db6	**0.347**	**0.220**	1	0.352	0.225
sym6	0.351	0.226	2	0.351	0.226
coif5	0.349	0.222	**3**	**0.347**	**0.220**
bior4.4	0.350	0.222	4	0.348	0.222
haar	0.349	0.221	5	0.348	0.220

Table 3: Average MSE comparison of ALW-enhanced CycleNet across wavelet families (left) and decomposition levels using the db6 wavelet (right). Best results are in **bold**.

stationarity beyond certain horizons where older patterns become irrelevant due to concept drift. Crucially, ALW still substantially mitigates performance degradation compared to fixed-window approaches, achieving 29.6% and 70.9% lower MSE at $L = 4320$ on ETTh1 and ETTh2, while simultaneously compressing inputs to $H = 256$. This demonstrates ALW's dual efficacy in enhancing predictive accuracy and improving resilience to long input sequences, with dataset characteristics and representational constraints ultimately determining the practical upper bound of useful historical information.

Analysis of Wavelet Hyperparameters We conducted a hyperparameter study to determine the optimal wavelet basis and decomposition level (J) for the ALW framework. First, we evaluated several common wavelet families (Table 3, left). db6 consistently outperforms sym6, coif5, bior4.4 and haar on both the ETTm1 and Weather datasets. db6 achieves optimal performance due to its balanced trade-off between its six vanishing moments and a 12-coefficient filter length. This provides sufficient smoothness to capture complex temporal patterns while maintaining compact support, which limits boundary artifacts. Consequently, we fixed db6 as our wavelet basis and proceeded to identify the best decomposition level. Our analysis (Table 3, right) shows that $J = 3$ yields the lowest MSE: lower levels ($J < 3$) produce coarse decompositions that fail to disentangle frequency bands, whereas higher levels ($J > 3$) generate too-short sub-sequences prone to noise and raise computational cost. Therefore, a three-level decomposition delivers the optimal balance between feature granularity and signal integrity.

Efficiency Analysis

To evaluate the interplay between performance gains and computational overhead introduced by the ALW framework, we analyzed model parameter count, Multiply-Accumulate Operations (MACs), training time per iteration, and MSE on the ETTh1 dataset ($F = 720$). Table 4 compares original backbone models ($L = 96$), backbones with extended fixed lookback (+lookback, $L = 512$), and backbones enhanced with ALW (+ALW, initial input $L = 512$, mapped to $H = 256$ for the backbone). Experimental results demonstrate that merely elongating the fixed lookback window does not uniformly improve MSE across all models, whereas

Model	Parameters	MACs	Train.Time	MSE
TQNet	981.75K	211.03M	5.3ms	0.487
+lookback	2.21M	258.74M	5.8ms	0.489
+ALW	1.54M	284.27M	15.2ms	0.446
CycleNet	419.19K	93.59M	3.8ms	0.520
+lookback	632.18K	141.30M	4.3ms	0.473
+ALW	749.54K	166.82M	13.5ms	0.448
iTransformer	304.72K	106.87M	10.7ms	0.503
+lookback	357.97K	125.62M	12.1ms	0.553
+ALW	573.63K	168.97M	22.5ms	0.486
DLinear	139.68K	30.99M	3.7ms	0.513
+lookback	738.72K	165.27M	4.2ms	0.472
+ALW	618.51K	137.52M	15.8ms	0.465

Table 4: Static and runtime metrics when ALW is applied to backbone models on the ETTh1 dataset with prediction length $F = 720$.

ALW consistently enhances prediction accuracy, highlighting its performance optimization capabilities. The computational overhead introduced by the ALW module is relatively low. Results show an average increase in training time of approximately 10ms, which is generally acceptable for long-term series forecasting tasks without stringent real-time requirements. ALW's impact on computational resources varies with model structure. For models whose complexity correlates strongly with input length, ALW effectively reduces parameter count and MACs through dimensionality reduction. For models weakly correlated with input length, there is a slight increase in these metrics due to the module's inherent computation. Nevertheless, compared to simply extending fixed lookback, ALW achieves improved accuracy with a superior performance-to-cost ratio. Furthermore, ALW's adaptive lookback window selection alleviates the need for hyperparameter tuning for optimal lookback length. In practice, this significantly diminishes the expensive and time-consuming search process traditionally required, enhancing the framework's practicality.

Conclusion

This paper addresses the critical limitations of fixed lookback windows in time series forecasting, which struggle to capture dynamic temporal dependencies and multi-scale patterns. We propose ALW, a wavelet transform-driven multi-scale adaptive framework. Its core integrates a wavelet-based Multi-scale Decomposition Module with a differentiable Adaptive Lookback Window Learning Module that leverages attention, backward accumulation, and soft truncation to determine instance-specific historical windows per scale. Subsequently, a Weighted Reconstruction Module effectively fuses this multi-scale information into a high-quality input for prediction. Extensive experiments validate ALW as an efficient, universal, and data-driven plug-and-play solution for long-term series forecasting. Future work will explore extensions to spatio-temporal forecasting.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62472332), and the Hubei Provincial International Science and Technology Cooperation Project (No.2024EHA031).

References

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 65–74.

Cao, W.; Wang, D.; Li, J.; Zhou, H.; Li, L.; and Li, Y. 2018. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Advances in Neural Information Processing Systems*, volume 31, 6776–6786.

Farlie, D. 1964. Prediction and Regulation by Linear Least-Square Methods. *Journal of the Operational Research Society*, 15(4): 410–411.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Huang, C.; Bai, C.; Chan, S.; Zhang, J.; and Wu, Y. Q. 2023. MGTCF: multi-generator tropical cyclone forecasting with heterogeneous meteorological data. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, volume 37, 5096–5104.

Jiang, J.; Han, C.; Zhao, W. X.; and Wang, J. 2023. PDFormer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the Thirty-Seventh AAAI conference on artificial intelligence*, volume 37, 4365–4373.

John, D.; Binnewies, S.; and Stantic, B. 2025. Dynamic Optimisation of Window Sizes for Enhanced Time-Series Forecasting. *Preprints*.

John, D. L.; Binnewies, S.; and Stantic, B. 2024. Identifying Optimal Window Size Configurations for Big Data Time Series Forecasting. In *Proceedings of the 2024 IEEE International Conference on Big Data*, 5138–5146.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *The third International Conference on Learning Representations*.

Koparanov, K. A.; Georgiev, K. K.; and Shterev, V. A. 2020. Lookback period, epochs and hidden states effect on time series prediction using a LSTM based neural network. In *Proceedings of the 28th National Conference with International Participation*, 61–64.

Li, S.; Cui, Y.; Li, L.; Yang, W.; Zhang, F.; and Zhou, X. 2024. ST-ABC: Spatio-Temporal Attention-Based Convolutional Network for Multi-Scale Lane-Level Traffic Prediction. In *Proceedings of the 40th IEEE International Conference on Data Engineering*, 1185–1198. IEEE.

Lin, S.; Chen, H.; Wu, H.; Qiu, C.; and Lin, W. 2025. Temporal Query Network for Efficient Multivariate Time Series Forecasting. In *International Conference on Machine Learning*. PMLR.

Lin, S.; Lin, W.; Hu, X.; Wu, W.; Mo, R.; and Zhong, H. 2024. CycleNet: Enhancing Time Series Forecasting through Modeling Periodic Patterns. *Advances in Neural Information Processing Systems*, 37: 106315–106345.

Liu, M.; Zeng, A.; Chen, M.; Xu, Z.; Lai, Q.; Ma, L.; and Xu, Q. 2022a. SCINet: Time series modeling and forecasting with sample convolution and interaction. In *Advances in Neural Information Processing Systems*, volume 35, 5816–5828.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022b. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 9881–9893.

Luo, Y.; Gu, Z.; Zhou, S.; Xiong, Y.; and Gao, X. 2023. Meteorology-Assisted Spatio-Temporal Graph Network for Uncivilized Urban Event Prediction. In Chen, G.; Khan, L.; Gao, X.; Qiu, M.; Pedrycz, W.; and Wu, X., eds., *Proceedings of the IEEE International Conference on Data Mining*, 468–477. IEEE.

Makridakis, S.; and Hibon, M. 1997. ARMA models and the Box–Jenkins methodology. *Journal of forecasting*, 16(3): 147–163.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32, 8024–8035.

Qiu, X.; Wu, X.; Lin, Y.; Guo, C.; Hu, J.; and Yang, B. 2025. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 1185–1196.

Tong, S.; and Yuan, J. 2025. Efficiently Enhancing Long-term Series Forecasting via Ultra-long Lookback Windows. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, 20912–20920.

Uremovic, N.; Bizjak, M.; Sukic, P.; Stumberger, G.; Zalik, B.; and Lukac, N. 2023. A New Framework for Multivariate Time Series Forecasting in Energy Management System. *IEEE Trans. Smart Grid*, 14(4): 2934–2947.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A. W.;

and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. In *The ninth ISCA Speech Synthesis Workshop*, 125.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 5998–6008.

Wang, H.; Peng, J.; Huang, F.; Wang, J.; Chen, J.; and Xiao, Y. 2023. MICN: Multi-scale Local and Global Context Modeling for Long-term Series Forecasting. In *The Eleventh International Conference on Learning Representations*.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and ZHOU, J. 2024. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.

Watson, M. W. 1994. Vector autoregressions and cointegration. *Handbook of econometrics*, 4: 2843–2915.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 22419–22430.

Xu, Z.; Zeng, A.; and Xu, Q. 2024. FITS: Modeling Time Series with $10k$ Parameters. In *The Twelfth International Conference on Learning Representations*.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are Transformers Effective for Time Series Forecasting? In Williams, B.; Chen, Y.; and Neville, J., eds., *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, volume 37, 11121–11128.

Zhang, Z.; Pham, T. D.; An, Y.; Doan, N. P.; Alsharari, M.; Tran, V.-H.; Hoang, A.-T.; Vandierendonck, H.; and Mai, S. T. 2025. WaveletMixer: a multi-resolution wavelets based MLP-mixer for multivariate long-term time series forecasting. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, 22741–22749.

Zhong, Y.; Wang, J.; Rao, J.; Xu, J.; and Wu, S. 2024. A Novel Series-Concatenation Hybrid Prediction Model of Energy Consumption in Hot Strip Roughing Process With Multi-Step Rolling. *IEEE Trans Autom. Sci. Eng.*, 21(3): 4585–4598.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *International Conference on Machine Learning*, volume 162, 27268–27286.