M^{2}FMoE: Multi-Resolution Multi-View Frequency Mixture-of-Experts for Extreme-Adaptive Time Series Forecasting

Yaohui Huang, Runmin Zou, Yun Wang*, Laeeq Aslam, Ruipeng Dong

School of Automation, Central South University, Changsha, China
{yaohuihuang, rmzou, wangyun19, laeeq_aslam, darol22}@csu.edu.cn

Abstract

Forecasting time series with extreme events is critical yet challenging due to their high variance, irregular dynamics, and sparse but high-impact nature. While existing methods excel in modeling dominant regular patterns, their performance degrades significantly during extreme events, constituting the primary source of forecasting errors in real-world applications. Although some approaches incorporate auxiliary signals to improve performance, they still fail to capture extreme events' complex temporal dynamics. To address these limitations, we propose M^{2}FMoE, an extreme-adaptive forecasting model that learns both regular and extreme patterns through multi-resolution and multi-view frequency modeling. It comprises three modules: (1) a multi-view frequency mixture-of-experts module assigns experts to distinct spectral bands in Fourier and Wavelet domains, with cross-view shared band splitter aligning frequency partitions and enabling inter-expert collaboration to capture both dominant and rare fluctuations; (2) a multi-resolution adaptive fusion module that hierarchically aggregates frequency features from coarse to fine resolutions, enhancing sensitivity to both short-term variations and sudden changes; (3) a temporal gating integration module that dynamically balances long-term trends and short-term frequency-aware features, improving adaptability to both regular and extreme temporal patterns. Experiments on real-world hydrological datasets with extreme patterns demonstrate that M^{2}FMoE outperforms state-of-the-art baselines without requiring extreme-event labels.

Code — https://github.com/Yaohui-Huang/M2FMoE

Introduction

Time series forecasting is vital for decision-making across various real-world systems, including energy, transportation, and environmental monitoring (Jin et al. 2024; Wang et al. 2024b). Among these, hydrological forecasting is particularly difficult due to extreme events like flash floods, heavy rainfall, and sudden water level rises (Lavers, Pappenberger, and Zsoter 2014). These events are rare, abrupt, and high variance, often causing significant deviations from regular temporal patterns (Camps-Valls et al. 2025). Despite their importance for risk management, forecasting such extremes remains one of the most challenging problems in time series modeling (Li, Xu, and Anastasiu 2024).

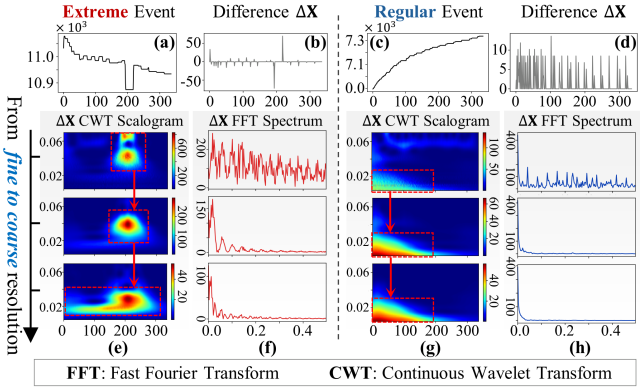

Figure 1: Comparison of frequency spectra between regular and extreme events.

Classical statistical models often fail under extreme or non-stationary conditions (Zhang 2003). Recent deep learning advancements offer enhanced flexibility in modeling intricate temporal dependencies (Wen et al. 2023; Wang et al. 2024b). However, these models typically emphasize capturing dominant patterns such as periodic trends, smooth transitions, and local correlations, resulting in the inadequate representation of rare, high-impact extreme events. Consequently, forecasting models tend to perform well under regular conditions but struggle to accurately represent these infrequent but critical dynamics. This limitation is especially pronounced in hydrological forecasting, where systems are highly sensitive to abrupt shifts, such as sudden heavy rainfall or rapid runoff (Li and Anastasiu 2025). Inaccurate predictions in such scenarios may lead to delayed warnings and severe consequences like widespread flooding. These challenges highlight the urgent need for forecasting models that can accurately capture both regular trends and extreme deviations within a unified framework.

Frequency-domain representations provide a promising way to decompose temporal dynamics into spectral components, facilitating models to separate high-frequency fluctuations from low-frequency trends (Ma et al. 2024; Liu 2025). The spectral characteristics of extreme and regular events

*Corresponding author.

are shown in **Fig.1**. As illustrated in **Fig.1(a)–(d)**, the differenced sequences $\Delta\mathbf{X}$ reveal clear contrasts between the two types of events. These differences become more pronounced in the wavelet domain (**Fig. 1(e), 1(g)**), where extreme events produce sharp, localized energy at fine resolutions. As the resolution becomes coarser, energy gradually shifts toward lower frequencies with reduced intensity, while the main structure of the event remains consistent across resolutions. In contrast, regular events exhibit smooth low-frequency dynamics, resulting in diffuse and uniform energy distributions at all resolutions. Similar patterns are observed in the Fourier domain (**Fig. 1(f), 1(h)**). Extreme sequences exhibit broad-spectrum, multi-peaked energy with slow spectral decay, whereas regular sequences concentrate energy within narrow low-frequency bands. These observations highlight the need for frequency-aware modeling to capture the varied spectral properties of temporal patterns. In particular, the results reveal ***frequency heterogeneity***, where different frequency bands contribute unequally to regular and extreme events. Accurately modeling such variation requires adaptive focus on informative frequencies, which is challenging within a single spectral domain. Fourier transforms provide accurate global frequency information but lack temporal resolution. In contrast, wavelet transforms offer time-frequency localization but suffer from reduced resolution at lower frequencies (Fei et al. 2025). Combining both views yields a more complete spectral representation that supports the modeling of both abrupt variations and long-term dependencies. Nevertheless, this dual-view strategy also introduces ***cross-view spectral misalignment***. Differences in basis functions and resolution cause the same signal to localize inconsistently across Fourier and Wavelet domains, resulting in cross-view incompatibility that undermines shared modeling.

To address this, we propose a Multi-resolution Multi-view Frequency Mixture-of-Experts (**M^{2}FMoE**) to model frequency-aware temporal dynamics under both regular and extreme conditions. Specifically, M^{2}FMoE first proposes a multi-view frequency mixture-of-experts (MFMoE) module to assigns specialized spectral experts to distinct frequency bands across both Fourier and Wavelet domains, thereby enabling selective specialization to handle diverse frequency characteristics. To ensure semantic coherence among experts and alleviate spectral misalignment, a cross-view shared band splitter (CSS) is integrated within MFMoE, aligning spectral boundaries across views. Furthermore, to capture temporal patterns at multiple resolutions, we introduce the multi-resolution adaptive fusion (MAF) module, which hierarchically aggregates features from coarse to fine frequency scales. Finally, a temporal gating integration (TGI) module adaptively fuses recent dynamics with long-range historical context via a learnable gating mechanism. Experiments on five hydrological datasets with extreme events demonstrate that M^{2}FMoE outperforms state-of-the-art methods without using auxiliary event labels.

Related Work

Time series forecasting has evolved from classical models like ARIMA, which are interpretable but limited by linearity and stationarity (Zhang 2003; Wang et al. 2024b), to deep learning methods that offer greater flexibility. Early deep learning approaches, including RNNs (Jia et al. 2024; Kong et al. 2025) and CNNs (Wu et al. 2023; Chen, Jiang, and Gel 2023), focused on local dependencies. Transformers (Liu et al. 2025a, 2024; Kim et al. 2024) then introduced self-attention for long-range structure, while recent MLP-based models (Lin et al. 2025, 2024; Liu et al. 2025b) offer efficient alternatives. Further advancements include GNNs (Jin et al. 2024, 2025a) and Mixture-of-Experts (MoE) models (Liu 2025; Shi et al. 2025) for nonlinear modeling, alongside multi-scale and multi-resolution representations (Wang et al. 2025, 2024a) for capturing varied temporal granularities. Frequency-based approaches have also emerged, utilizing Fourier transforms for global periodic structures and wavelet transforms for localized time-frequency representations (Ma et al. 2024; Fei et al. 2025). However, most existing models primarily target regular patterns, struggling with the irregular variations crucial for extreme events.

Extreme-adaptive forecasting targets time series with rare, abrupt, and high-impact changes such as floods or sudden surges in water levels, which require effective handling of rarity and volatility. Recent efforts span architectural designs and loss functions. Architecturally, models such as NEC+ (Li, Xu, and Anastasiu 2023a), VIE (Xiu et al. 2021), SADI (Liu et al. 2023), and SEED (Li, Xu, and Anastasiu 2023b) use multi-phase learning for non-stationary dynamics. Others, like MCANN (Li and Anastasiu 2025) and DAN (Li, Xu, and Anastasiu 2024), integrate priors or clustering for enhanced robustness. On the loss side, specialized objectives like EPL (Wang, Han, and Guo 2024), EVL (Ding et al. 2019), and GEVL (Zhang et al. 2021) leverage Extreme Value Theory or biases to emphasize tail behavior. Despite these advancements, current extreme-adaptive methods often neglect the joint modeling of frequency-aware patterns and resolution-specific dynamics, limiting their generalization across diverse temporal variations.

Preliminaries

Problem Statement Let $\mathbf{X} = \{X_1, X_2, \ldots, X_{T_{in}}\}$ denote a multivariate time series, where each $X_t \in \mathbb{R}^C$ represents a C-dimensional observation at time step t, and T_{in} is the input sequence length. The objective is to learn a forecasting model $\mathbb{F}(\cdot)$ that predicts the subsequent T_p future values, denoted as $\hat{\mathbf{X}} = \{\hat{X}_{T_{in}+1}, \ldots, \hat{X}_{T_{in}+T_p}\}$.

Discrete Fourier Transform (DFT) The discrete Fourier transform (DFT) decomposes a sequence into global sinusoidal components. For $\mathbf{X}$ of length T_{in}, its n-th frequency coefficient is:

$$\mathcal{F}_n = \sum_{t=1}^{T_{in}} X_t \cdot e^{-j2\pi nt/T_{in}}, \quad n \in \{0, 1, \ldots, T_{in}-1\}, \quad (1)$$

where $\mathcal{F}_n$ encodes periodicity but lacks temporal localization, limiting its applicability for non-stationary signals. FFT is implemented to efficiently compute $\mathcal{F}_n$.

Continuous Wavelet Transform (CWT) The CWT enables localized time-frequency analysis. For a signal $X(t)$,

the wavelet coefficient at scale a and position b is defined as:

$$\mathcal{W}(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} X(t)\, \psi^* \left(\frac{t-b}{a} \right) dt, \qquad (2)$$

where ψ^* is the complex conjugate of the mother wavelet ψ.

Methodology

As illustrated in **Fig. 2**, the proposed M^{2}FMoE comprises three modules: (1) an MFMoE module, (2) an MAF module, and (3) a TGI module. Each component is detailed below.

Hierarchical Temporal Segmentation As shown in **Fig. 2(a)**, the hierarchical temporal segmentation module extracts a recent segment $\mathbf{X}_r = \{X_{T_{in}-T_r+1}, \ldots, X_{T_{in}}\}$ to capture short-term dynamics, while the entire input sequence $\mathbf{X} = \{X_1, \ldots, X_{T_{in}}\}$ serves as the historical context to model long-term temporal patterns.

Multi-Resolution Sequence Generation To capture temporal dynamics at varying granularities, the recent segment $\mathbf{X}_r \in \mathbb{R}^{T_r \times C}$ is decomposed into a multi-resolution set via 1D smoothing convolutions:

$$\mathcal{S} = \left\{ \tilde{\mathbf{X}}_r^{(k)} = \text{SmoothConv}(\mathbf{X}_r, k) \mid k \in \mathcal{K} \right\}. \qquad (3)$$

For each resolution k (with $k_1 = 1$ retaining the original sequence), we compute the first-order difference of the transformed recent sequence to highlight local variations, i.e., $\Delta\mathbf{X}_r^{(k)} = \tilde{\mathbf{X}}_r^{(k)}[1:T_r] - \tilde{\mathbf{X}}_r^{(k)}[0:T_r-1]$. The resulting multi-resolution differences $\Delta\mathbf{X}_r^{(k)}$ disentangle coarse and fine dynamics, forming a diverse input set for subsequent frequency-aware modeling in the MFMoE module.

Temporal Embedding To encode temporal order, a fixed sinusoidal positional embedding is added as an auxiliary feature (Liu et al. 2024). Each time step is mapped to a combination of sine and cosine functions at varying frequencies:

$$\text{PE}(t, 2i) = \sin(t \cdot \omega_i), \quad \text{PE}(t, 2i+1) = \cos(t \cdot \omega_i), \quad (4)$$

where t is the time index and d the embedding dimension, i is the index of the embedding dimension, and $\omega_i = 1/10000^{2i/d}$. These embeddings are added to input features to provide position-aware inductive bias across time steps.

Multi-View Frequency Mixture-of-Experts Module

The recent segment of extreme time series is challenging due to the sparsity and volatility of extreme events. To address this, we shift the learning paradigm to the frequency domain via the MFMoE module, which comprises two expert branches: a Fourier-view and a Wavelet-view branch. A CSS is introduced to align frequency bands across both domains.

Cross-View Shared Band Splitter To capture multi-resolution temporal dynamics, we construct dual-view spectral representations using both Fourier and Wavelet transforms. However, a key challenge arises from the inherent differences in these views. The Fourier transform organizes spectral components on a uniform frequency axis, while the CWT uses scales that correspond nonlinearly to frequency. Consequently, aligning expert assignments between the two views is difficult, as the same frequency content can appear at different positions in each representation.

To ensure consistent expert specialization across both spectral views, we propose the *Cross-View Shared Band Splitter*. **Theorem 1** provides the theoretical basis for this module by formalizing the correspondence between frequency and wavelet scale. As shown in **Fig. 2(b)**, the splitter learns shared frequency boundaries $\{\beta_1, \beta_2, \ldots, \beta_{E-1}\}$ to divide the frequency range $[0, 1]$ into E bands. For the FFT view, these boundaries are directly scaled into frequency indices $\{\tilde{\beta}_1, \ldots, \tilde{\beta}_{E-1}\}$. For the CWT view, they are nonlinearly mapped into wavelet scales $\{\ddot{\beta}_1, \ldots, \ddot{\beta}_{E-1}\}$ using the inverse relationship from **Theorem 1**. This shared segmentation allows both views to decompose the input into semantically aligned sub-bands, ensuring experts operate on consistent spectral content.

Theorem 1 (Spectral Boundary Correspondence)**.** *Let f denote the normalized frequency, a is the scale in the CWT, and $\gamma = f_0/f_{\text{nyq}}$ is a wavelet-dependent constant. The mapping $a = \gamma/f$ establishes a one-to-one correspondence between frequency and scale boundaries (Mallat 2002), such that $f_{\max} \mapsto a_{\min} = \gamma/f_{\max}$ and $f_{\min} \mapsto a_{\max} = \gamma/f_{\min}$. Under this mapping, signal energy is conserved, satisfying $\int_{f_{\min}}^{f_{\max}} |\mathcal{F}(f)|^2 df \propto \iint_{a \in [a_{\min}, a_{\max}]} |\mathcal{W}(a,b)|^2 \frac{da}{a^2} db$.*

Fourier-View Expert Branch As illustrated in **Fig. 2(c)**, the Fourier-view expert branch is designed to extract frequency-aware representations by assigning specialized experts to distinct frequency bands. Using the shared boundaries $\{\tilde{\beta}_1, \ldots, \tilde{\beta}_{E-1}\}$, we divide the full frequency range into E non-overlapping intervals (i.e., $[0, \tilde{\beta}_1), [\tilde{\beta}_1, \tilde{\beta}_2), \ldots, [\tilde{\beta}_{E-1}, F]$), where $F = T_r/2 + 1$ is the number of frequency bins after real-valued FFT. Each expert e is responsible for modeling the frequency components within its assigned band, such as high-frequency, mid-frequency, and low-frequency patterns.

Given an input sequence $\Delta\mathbf{X}_r^{(k)} \in \mathbb{R}^{T_r \times C}$, we first perform per-channel standardization and apply the real FFT. Then, to isolate expert-specific frequency components, we define a set of binary masks $\tilde{\mathbb{I}}_e \in \mathbb{R}^{F \times C}$, each indicating the active sub-band for expert e. The masked spectrum for expert e is:

$$\mathcal{F}_e = \tilde{\mathbb{I}}_e \odot \mathcal{F}, \quad \tilde{\mathbb{I}}_e = \begin{cases} 1, & \text{if } f \in [\tilde{\beta}_{e-1}, \tilde{\beta}_e) \\ 0, & \text{otherwise} \end{cases}, \qquad (5)$$

where $\mathcal{F}$ represents the full spectrum generated by applying FFT to $\Delta\mathbf{X}_r^{(k)}$. To adaptively determine expert contributions, inspired by (Jin et al. 2025b), we employ a lightweight routing network. Specifically, we first compute the magnitude spectrum and average across channels, and then the summary vector is passed through a routing network $\tilde{\mathcal{G}}(\cdot)$ consisting of two linear layers with ReLU activation and softmax output to produce the expert routing weights:

$$\tilde{\mathbf{M}} = \frac{1}{C} \sum\nolimits_{c=1}^{C} |\mathcal{F}_e[c]|, \quad \boldsymbol{\alpha} = \text{Softmax}\big(\tilde{\mathcal{G}}(\tilde{\mathbf{M}})\big), \qquad (6)$$

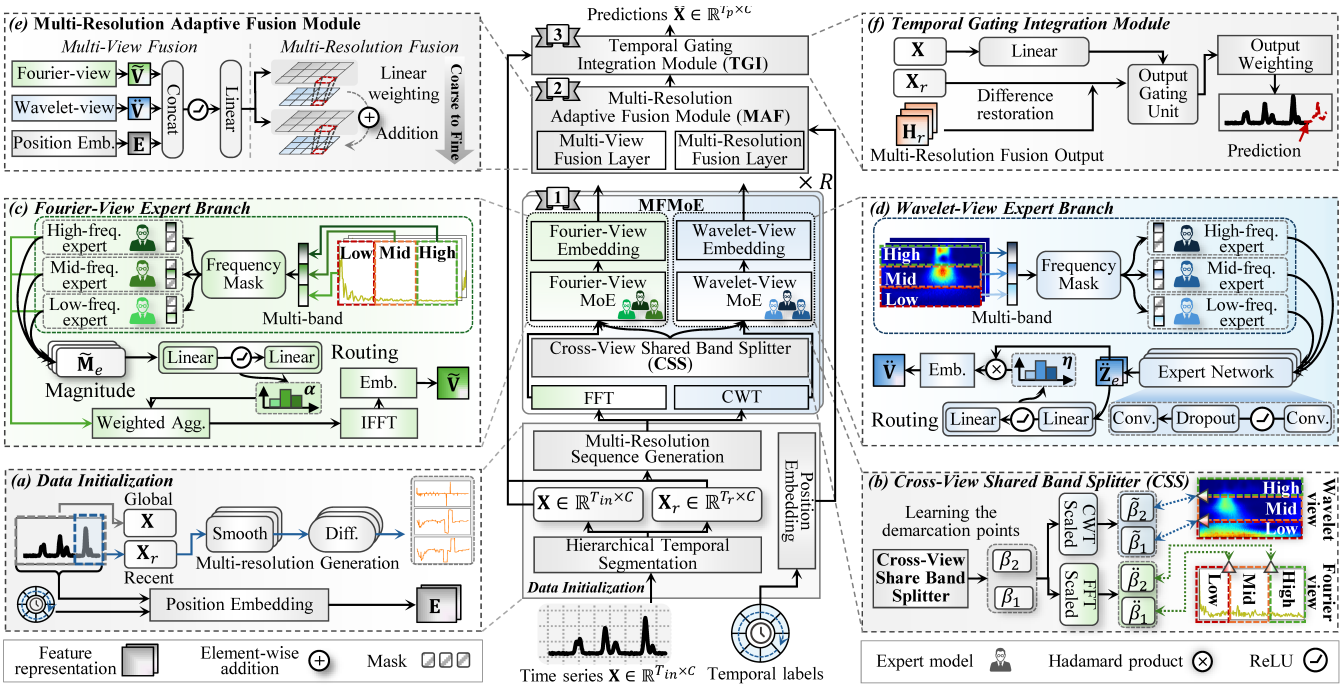

Figure 2: The proposed M^{2}FMoE with three experts per branch for capturing high-, mid-, and low-frequency patterns.

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_E]$ are the routing weights for each expert. $\tilde{\mathbf{M}} \in \mathbb{R}^F$ is the magnitude spectrum averaged across channels. The final output of the Fourier-view expert branch is obtained by aggregating expert-specific frequency components using routing weights α_e, followed by inverse FFT and a linear projection:

$$\tilde{\mathbf{V}} = \text{Linear}\big(\text{IFFT}(\sum\nolimits_{e=1}^{E} \alpha_e \cdot \mathcal{F}_e)\big), \tag{7}$$

where $\tilde{\mathbf{V}} \in \mathbb{R}^{T_p \times C}$ is the final output of the Fourier-view expert branch. This design enables dynamic selection of frequency bands and temporal adaptation based on input spectral statistics.

Wavelet-View Expert Branch As illustrated in **Fig. 2(d)**, the Wavelet-view expert branch captures temporally localized dynamics by operating on the CWT power spectrogram $\mathcal{P} = |\mathcal{W}(a,b)|^2 \in \mathbb{R}^{C \times S \times T_r}$, where S is the number of wavelet scales. The CWT is computed using the complex Gaussian wavelet, ensuring balanced localization in both time and frequency domains.

The shared frequency boundaries are first converted to scale indices $\{\ddot{\beta}_1, \ldots, \ddot{\beta}_{E-1}\}$ via the inverse mapping defined in **Theorem 1**. Each expert e is assigned a binary scale mask $\mathbb{I}_e \in \{0,1\}^S$, and its corresponding component is computed as:

$$\mathcal{P}_e = \ddot{\mathbb{I}}_e \odot \mathcal{P}, \quad \ddot{\mathbb{I}}_e = \begin{cases} 1, & \text{if } s \in [\ddot{\beta}_{e-1}, \ddot{\beta}_e), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Each expert network processes its masked input $\mathcal{P}_e$ using a convolutional block:

$$\ddot{\mathbf{Z}}_e = \mathbf{W}_{e,2} * \Big(\mathcal{D}\big(\text{ReLU}\left(\mathbf{W}_{e,1} * \mathcal{P}_e\right)\big)\Big), \tag{9}$$

where $*$ denotes the convolution operation, $\mathbf{W}_{e,1}, \mathbf{W}_{e,2}$ are convolution kernels, and the dropout layer $\mathcal{D}(\cdot)$ is applied after the activation to prevent overfitting.

To adaptively assign expert contributions, the power spectrogram $\mathcal{P}$ is first averaged over the channel dimension to obtain a global summary $\ddot{\mathbf{M}} \in \mathbb{R}^{S \times T_r}$. This matrix is then flattened and passed through a lightweight routing network $\ddot{\mathcal{G}}(\cdot)$, which consists of two linear layers with ReLU activation, followed by a softmax function to produce the expert weighting vector:

$$\ddot{\mathbf{M}} = \frac{1}{C}\sum\nolimits_{c=1}^{C} \mathcal{P}[c], \quad \boldsymbol{\eta} = \text{Softmax}\left(\ddot{\mathcal{G}}\left(\text{Flatten}(\ddot{\mathbf{M}})\right)\right), \tag{10}$$

where $\boldsymbol{\eta}$ represents the soft assignment weights over the E experts for a given input. The final output is obtained via gated aggregation of expert outputs, followed by two linear layers to project the result to the target shape:

$$\ddot{\mathbf{V}} = \mathbf{W}_{o,1}\left(\mathbf{W}_{o,2}\left(\text{Flatten}\left(\sum\nolimits_{e=1}^{E} \eta_e \cdot \ddot{\mathbf{Z}}_e\right)\right)^{\top}\right), \tag{11}$$

where $\mathbf{W}_{o,1}, \mathbf{W}_{o,2}$ are learnable weighting matrices, and $\ddot{\mathbf{V}} \in \mathbb{R}^{T_p \times C}$ is the final output of the Wavelet-view expert branch. $\eta_e \in \mathbb{R}$ is the gating weight for expert e.

Multi-Resolution Adaptive Fusion Module

The MAF module consists of two key phases: (1) a multi-view fusion phase and (2) a multi-resolution fusion phase, as illustrated in **Fig. 2(e)**.

In the **multi-view fusion** phase, the temporal outputs from the Fourier and Wavelet expert branches, denoted as

$\{\tilde{\mathbf{V}}, \ddot{\mathbf{V}}\} \in \mathbb{R}^{T_p \times C}$, are concatenated along the channel axis with temporal encoding $\mathbf{E} \in \mathbb{R}^{T_p \times 2}$. Then, the fused representation is processed by a stacked projection block with two linear layers and batch normalization:

$$\mathbf{H}_u^{(i)} = \mathbf{W}_{u,2} \cdot \mathcal{D}\left(\text{ReLU}\left(\text{BN}\left(\mathbf{W}_{u,1}[\tilde{\mathbf{V}}^{(i)}; \ddot{\mathbf{V}}^{(i)}; \mathbf{E}]^\top\right)\right)\right), \tag{12}$$

where $\mathbf{W}_{u,1} \in \mathbb{R}^{H' \times (2C+2)}$ and $\mathbf{W}_{u,2} \in \mathbb{R}^{C \times H'}$ are learnable weights; H' is the hidden dimension; $\text{BN}(\cdot)$ denotes batch normalization; and $[\cdot;\cdot]$ indicates channel-wise concatenation. The output $\mathbf{H}_u^{(i)} \in \mathbb{R}^{T_p \times C}$ represents the unified feature at the i-th resolution, with $i \in \{1, 2, \ldots, R\}$, where R is the total number of resolutions.

In the **multi-resolution fusion** phase, representations from different resolutions are projected into a shared space and combined via additive accumulation:

$$\mathbf{H}_r = \sum\nolimits_{i=1}^{R} \text{Linear}_i(\mathbf{H}_u^{(i)}) \in \mathbb{R}^{T_p \times C}, \tag{13}$$

where $\text{Linear}_i(\cdot)$ is a resolution-specific linear transformation. Since all frequency-view representations are learned from differenced sequences, the final fused output is shifted by adding back the last observed input slice to restore the original value space. This process enables coarse-to-fine feature refinement and enhances the integration of multi-scale temporal dynamics.

Temporal Gating Integration Module

The TGI module adaptively combines the recent prediction and historical scene representation to produce the final output, as illustrated in **Fig. 2(f)**. To adaptively integrate the recent prediction and the historical scene representation, a gating mechanism is applied. Let $\mathbf{H}_r \in \mathbb{R}^{T_p \times C}$ denote the output from the multi-resolution fusion module, and $\mathbf{H}_h \in \mathbb{R}^{T_p \times C}$ be the transformed embedding of the historical input, obtained via a linear projection $\mathbf{H}_h \leftarrow \mathbf{W}_g \mathbf{X}$ with $\mathbf{W}_g \in \mathbb{R}^{T_{in} \times T_p}$. The gating coefficient is computed as:

$$\mathbf{G} = \sigma\left(\text{Linear}\left([\mathbf{H}_r; \mathbf{H}_h]\right)\right), \tag{14}$$

$$\hat{\mathbf{X}} = \mathbf{G} \odot \mathbf{H}_r + (1 - \mathbf{G}) \odot \mathbf{H}_h, \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid activation. $\hat{\mathbf{X}} \in \mathbb{R}^{T_p \times C}$ is the final output of the model.

Optimization Objective

The overall objective of M^{2}FMoE consists of three components. The primary term is the forecasting loss $\mathcal{L}_{\text{pred}}$, measured by Mean Squared Error (MSE). To promote diverse specialization within each branch and ensure consistency across branches, we introduce a regularization term comprising the expert diversity loss $\mathcal{L}_{\text{div}}$ and expert consistency loss $\mathcal{L}_{\text{cons}}$:

$$\mathcal{L}_{\text{div}} = \sqrt{\frac{1}{E}\sum\nolimits_{e=1}^{E}\left(\|\mathbf{Z}_e\|_2 - \frac{1}{E}\sum\nolimits_{j=1}^{E}\|\mathbf{Z}_j\|_2\right)^2}, \tag{16}$$

$$\mathcal{L}_{\text{cons}} = \frac{1}{E}\sum\nolimits_{e=1}^{E}\left(1 - cossim\left(\tilde{\mathbf{Z}}_e,\ \ddot{\mathbf{Z}}_e\right)\right), \tag{17}$$

where $\mathbf{Z}_e \in \{\tilde{\mathbf{Z}}_e, \ddot{\mathbf{Z}}_e\}$ is the output of the e-th expert in either the Fourier-view or Wavelet-view expert branch, $\|\cdot\|_2$ denotes the ℓ_2 norm, and $cossim(\cdot,\cdot)$ denotes the cosine similarity function. Here, $\tilde{\mathbf{Z}}_e$ indicates the inverse FFT of the masked frequency component $\mathcal{F}_e$ for the e-th expert, and $\ddot{\mathbf{Z}}_e$ is the output of the e-th expert in the Wavelet-view expert branch. The expert diversity loss $\mathcal{L}_{\text{div}}$ encourages the outputs of different experts to be diverse, while the expert consistency loss $\mathcal{L}_{\text{cons}}$ encourages the outputs of the same expert in different branches to be consistent. Finally, the overall optimization objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda\mathcal{L}_{\text{div}} + \mu\mathcal{L}_{\text{cons}}, \tag{18}$$

where λ and μ are hyperparameters that control the trade-off between the forecasting loss and the regularization terms.

Experiments

Experimental Settings

Datasets The experiment uses five public datasets containing hourly water level records from reservoirs in Santa Clara County, California. The datasets include Almaden, Coyote, Lexington, Stevens Creek, and Vasona, spanning the period from 1991 to 2019. Following the experimental protocol of (Li and Anastasiu 2025), the training and validation sets are randomly sampled from data between January 1991 and June 2018. The forecasting task targets the period from July 2018 to June 2019. To alleviate the data imbalance, we employed the same oversampling strategy as described in (Li and Anastasiu 2025).

Benchmarks To ensure a comprehensive evaluation, nine representative state-of-the-art baselines are selected: attention-based models (CATS (Kim et al. 2024), TQNet (Lin et al. 2025), iTransformer (iTrans.) (Liu et al. 2024)), frequency-domain models (FreqMoE (Liu 2025), Umixer (Ma et al. 2024)), linear-based models (KAN (Liu et al. 2025b), CycleNet (Lin et al. 2024)), and two extreme-enhanced methods that leverage event labels (DAN (Li, Xu, and Anastasiu 2024), MCANN (Li and Anastasiu 2025)).

Implementation Details For fair evaluation, the optimal configurations from the official implementations of MCANN and DAN are adopted. Following the experimental protocol in (Li and Anastasiu 2025), the prediction horizons are set to 8 and 72 hours, with a look-back window of 360 hours (i.e., 15 days). For other methods, standard baseline practices are followed: all datasets are normalized using z-score normalization, and denormalization is applied during evaluation to ensure predictions are in the original scale. The models are trained using the Adam (Kingma and Ba 2015) optimizer with a batch size of 48. Following the official protocol, evaluation is performed using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Main Results

Comparison with Benchmarks As shown in **Table 1**, we compare the proposed M^{2}FMoE model with nine state-of-the-art baselines on five reservoirs with prediction horizons of 8 and 72 hours. The results demonstrate that M^{2}FMoE

Data	Metrics	Horizon	*without extreme lables*								*with extreme lables*	
			M^{2}FMoE	**CATS**	**CycleNet**	**FreqMoE**	**iTrans.**	**KAN**	**TQNet**	**Umixer**	**DAN**	**MCANN**
Almaden	RMSE	8	**7.990**	16.087	17.754	14.729	32.127	18.934	18.023	18.658	37.857	8.447
	MAPE		**0.002**	0.006	0.007	0.005	0.017	0.009	0.010	0.007	0.021	**0.002**
	RMSE	72	**54.120**	57.916	61.379	63.038	65.325	70.181	59.427	64.816	66.597	56.840
	MAPE		**0.015**	**0.015**	0.019	0.017	0.025	0.033	0.018	0.018	0.025	**0.015**
Coyote	RMSE	8	**48.797**	110.849	113.706	593.141	372.523	116.398	103.521	174.892	505.941	86.829
	MAPE		**0.002**	0.004	0.003	0.018	0.022	0.005	0.003	0.005	0.025	**0.002**
	RMSE	72	**449.944**	509.077	528.962	855.096	673.853	587.132	504.606	566.429	829.623	559.747
	MAPE		**0.012**	**0.012**	**0.012**	0.025	0.029	0.021	**0.012**	0.013	0.042	**0.012**
Lexington	RMSE	8	**251.957**	618.991	463.293	386.995	690.426	429.054	400.991	466.669	476.936	252.965
	MAPE		0.004	0.011	0.011	0.006	0.041	0.008	0.013	0.008	0.015	**0.003**
	RMSE	72	**772.836**	906.531	865.092	1003.818	960.652	956.134	860.456	829.541	908.308	778.023
	MAPE		**0.014**	0.020	0.021	0.018	0.048	0.020	0.025	0.018	0.024	0.015
Stevens Creek	RMSE	8	**10.559**	18.500	28.400	80.937	48.876	25.672	24.475	37.654	24.319	12.130
	MAPE		**0.002**	0.004	0.005	0.017	0.010	0.005	0.006	0.007	0.011	**0.002**
	RMSE	72	**76.939**	82.739	94.578	117.282	106.606	94.034	89.265	141.505	82.794	81.084
	MAPE		0.014	**0.011**	0.014	0.025	0.017	0.015	0.012	0.017	0.020	**0.011**
Vasona	RMSE	8	**5.129**	6.913	7.903	14.318	12.179	11.308	7.741	9.299	9.562	5.353
	MAPE		**0.004**	0.007	0.007	0.020	0.013	0.019	0.007	0.009	0.012	**0.004**
	RMSE	72	19.571	20.381	20.713	20.740	21.534	21.605	20.173	23.718	20.542	**18.634**
	MAPE		0.021	0.021	0.021	0.027	0.023	0.027	0.020	0.023	0.023	**0.019**
Average Rank / Significance			**1.4**	3.7 / ⋆	4.9 / ∗	7.3 / ∗	8.5 / ∗	7.0 / ∗	4.4 / ∗	6.3 / ∗	7.9 / ∗	1.7 / ⋆

Table 1: Performance comparison on five reservoirs with predicted length as {8, 72} hours. ∗: both metrics are statistically significant ($p < 0.05$, Wilcoxon signed-rank test); ⋆: indicates significance in RMSE. Best results are **bold**, second-best underlined.

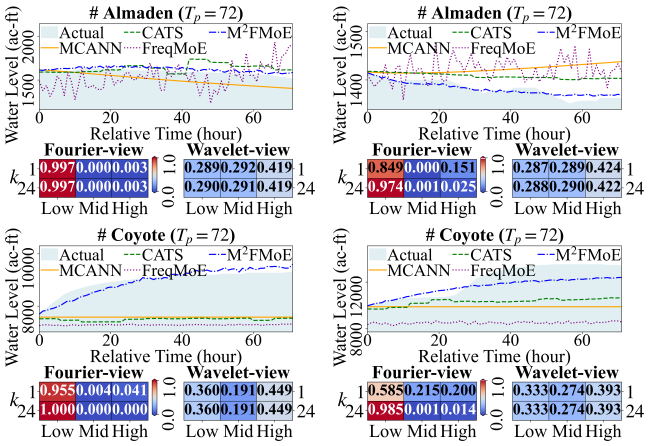

Figure 3: Prediction results and expert weights of M^{2}FMoE.

achieves the best average rank across all datasets and prediction horizons, outperforming all baselines in most cases. The improvements in RMSE are statistically significant on all reservoirs according to the Wilcoxon signed-rank test. Specifically, M^{2}FMoE achieves the average improvement of 22.30% over the best baseline without extreme labels and the maximum RMSE improvement of 52.86% on the Coyote dataset with a prediction horizon of 8 hours. Compared to the baselines with extreme labels, M^{2}FMoE also achieves competitive performance, with an average RMSE improvement of 9.19% across all settings, and a maximum improvement of 43.8% on the Coyote dataset with a prediction horizon of 8 hours. These results indicate that the proposed M^{2}FMoE model effectively captures the complex temporal dynamics of reservoir water levels, demonstrating its superiority over existing methods. **Fig. 3** presents the prediction results and expert weights of M^{2}FMoE, using three experts under two temporal resolutions (k=1 and k=24). The results show that Fourier-view experts primarily capture low-frequency trends, while Wavelet-view experts provide complementary high-frequency details. The adaptive weighting mechanism dynamically adjusts expert contributions based on input characteristics, improving M^{2}FMoE's performance on both regular and extreme events.

Ablation Studies We conduct ablation studies to evaluate the effectiveness of each component in the proposed M^{2}FMoE model. The ablation experiments are performed on the five reservoirs with a prediction horizon of 72 hours. The results are summarized in **Table 2**. The ablation studies include the following variants: (1) *w/o-WaveletView*: removes the Wavelet-view expert branch, (2) *w/o-FourierView*: removes the Fourier-view expert branch, (3) *w/o-*$\mathcal{L}_{\text{div}}$ & $\mathcal{L}_{\text{cons}}$: removes the expert diversity and consistency losses, (4) *w/o-Multi-Res*: utilizes the single-resolution and removes the multi-resolution fusion module,

Model	Almaden		Coyote		Lexington		Stevens Creek		Vasona	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
M^{2}FMoE	**54.120**	**0.015**	449.944	**0.012**	**772.836**	**0.014**	**76.939**	0.014	19.571	**0.021**
w/o-WaveletView	57.697	0.016	555.641	0.014	827.392	0.016	87.194	0.016	19.836	0.022
w/o-FourierView	59.035	0.020	558.262	0.014	870.735	0.017	85.022	0.017	19.813	0.024
w/o-$\mathcal{L}_{\text{div}}$ & $\mathcal{L}_{\text{cons}}$	54.950	0.017	**448.150**	0.013	826.408	0.015	77.293	**0.012**	20.080	**0.021**
w/o-Multi-Res	59.479	0.017	483.223	0.013	855.236	0.017	85.004	0.017	**19.508**	0.022
w/o-CSS	55.575	0.017	541.594	0.013	916.925	0.016	85.997	0.019	20.001	**0.021**

Table 2: Ablation study results on the five reservoirs with a prediction horizon of 72 hours.

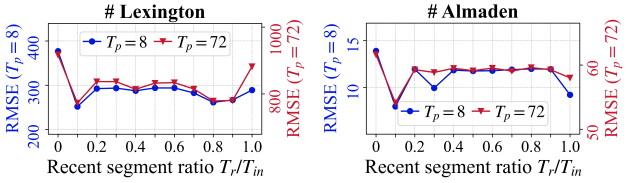

Figure 4: Impact of the length of recent segment T_r/T_{in}.

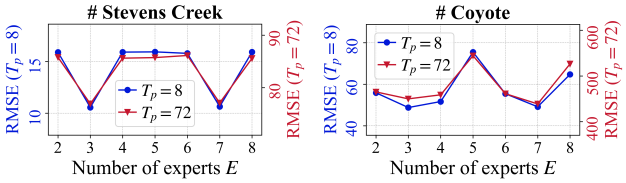

Figure 5: Impact of the number of experts E.

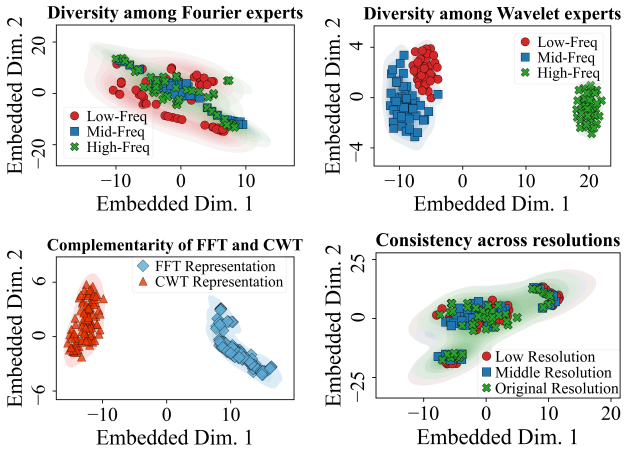

Figure 6: The t-SNE visualization of feature representations.

and (5) *w/o-CSS*: replace the cross-view shared band splitter with a uniform band splitter. We observe that the full model achieves the optimal performance across almost all datasets, demonstrating the effectiveness of the proposed multi-view and multi-resolution fusion strategy.

Impact of the Recent Segment Length The length of the recent segment critically affects the model's ability to capture extreme events. As shown in **Fig. 4**, reducing its length appropriately improves prediction accuracy by emphasizing relevant information. However, removing it entirely causes a significant performance drop, while overly long segments introduce noise and weaken the model's focus on extremes.

Impact of the Number of Experts We further examined the impact of expert count in M^{2}FMoE. As shown in **Fig. 5**, increasing the number of experts may improve pattern diversity but can also introduce noise and overfitting, leading to performance instability. Results suggest that using a moderate number (e.g., 3 or 4) yields the best predictive accuracy.

Visualization Analysis To better interpret the feature representations learned by M^{2}FMoE, the t-SNE (Maaten and Hinton 2008) is employed to visualize expert embeddings trained on the Almaden dataset using three spectral experts corresponding to high-, mid-, and low-frequency bands, as shown in **Fig. 6**. The visualizations reveal the following insights: (1) In the Fourier view, low-frequency features form a well-separated cluster from mid- and high-frequency features, indicating its effectiveness in capturing global trends. (2) The Wavelet view exhibits clearer separation between mid- and high-frequency features, suggesting superior sensitivity to localized, sparse patterns. (3) The cross-view distribution highlights the complementary nature of the two spectral views, with distinct clustering structures in each domain. (4) The cross-resolution view demonstrates that the multi-resolution fusion module maintains consistency while also capturing local variations. These results validate the proposed multi-view and multi-resolution strategy for effectively modeling diverse temporal patterns.

Conclusion

This study proposes M^{2}FMoE, an extreme-adaptive time series forecasting model that leverages multi-view frequency learning and multi-resolution fusion to capture both global trends and local extreme variations. Specifically, M^{2}FMoE employs specialized Fourier- and Wavelet-based experts to extract multi-frequency representations, while a multi-resolution fusion module progressively integrates temporal dependencies across resolutions. The model is optimized using forecasting, diversity, and consistency losses to promote adaptive and complementary expert behavior. Experiments on five reservoir datasets show that M^{2}FMoE outperforms state-of-the-art methods without using extreme event labels.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62376289, in part by the Natural Science Foundation of Hunan Province, China under Grant 2024JJ4069, and in part supported by the Fundamental Research Funds for the Central Universities of Central South University.

References

Camps-Valls, G.; Fernández-Torres, M.-Á.; Cohrs, K.-H.; Höhl, A.; Castelletti, A.; Pacal, A.; Robin, C.; Martinuzzi, F.; Papoutsis, I.; Prapas, I.; et al. 2025. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, 16(1): 1919.

Chen, Y.; Jiang, T.; and Gel, Y. R. 2023. H^2-Nets: Hyper-hodge Convolutional Neural Networks for Time-Series Forecasting. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, volume 14173, 271–289. Turin, Italy: Springer.

Ding, D.; Zhang, M.; Pan, X.; Yang, M.; and He, X. 2019. Modeling Extreme Events in Time Series Prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1114–1122. Anchorage, AK, USA: Association for Computing Machinery.

Fei, J.; Yi, K.; Fan, W.; Zhang, Q.; and Niu, Z. 2025. Amplifier: Bringing Attention to Neglected Low-Energy Components in Time Series Forecasting. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, volume 39, 11645–11653.

Jia, Y.; Lin, Y.; Yu, J.; Wang, S.; Liu, T.; and Wan, H. 2024. PGN: The RNN's New Successor is Effective for Long-Range Time Series Forecasting. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*. Vancouver, BC, Canada.

Jin, M.; Koh, H. Y.; Wen, Q.; Zambon, D.; Alippi, C.; Webb, G. I.; King, I.; and Pan, S. 2024. A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10466–10485.

Jin, M.; Shi, G.; Li, Y.-F.; Xiong, B.; Zhou, T.; Salim, F. D.; Zhao, L.; Wu, L.; Wen, Q.; and Pan, S. 2025a. Towards Expressive Spectral-Temporal Graph Neural Networks for Time Series Forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6): 4926–4939.

Jin, P.; Zhu, B.; Yuan, L.; and Yan, S. 2025b. MOH: Multi-head attention as mixture-of-head attention. In *ICML*.

Kim, D.; Park, J.; Lee, J.; and Kim, H. 2024. Are self-attentions effective for time series forecasting? In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, 114180–114209.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, CA, USA.

Kong, Y.; Wang, Z.; Nie, Y.; Zhou, T.; Zohren, S.; Liang, Y.; Sun, P.; and Wen, Q. 2025. Unlocking the Power of LSTM for Long Term Time Series Forecasting. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 11968–11976. Philadelphia, PA, USA.

Lavers, D. A.; Pappenberger, F.; and Zsoter, E. 2014. Extending medium-range predictability of extreme hydrological events in Europe. *Nature Communications*, 5(1): 5382.

Li, Y.; and Anastasiu, D. C. 2025. MC-ANN: A Mixture Clustering-Based Attention Neural Network for Time Series Forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8): 6888–6899.

Li, Y.; Xu, J.; and Anastasiu, D. 2024. Learning from polar representation: An extreme-adaptive model for long-term time series forecasting. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 1, 171–179. Vancouver, Canada.

Li, Y.; Xu, J.; and Anastasiu, D. C. 2023a. An Extreme-Adaptive Time Series Prediction Model Based on Probability-Enhanced LSTM Neural Networks. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 7, 8684–8691. Washington, DC, USA.

Li, Y.; Xu, J.; and Anastasiu, D. C. 2023b. SEED: An Effective Model for Highly-Skewed Streamflow Time Series Data Forecasting. In *Proceedings of the 2023 IEEE International Conference on Big Data*, 728–737. Sorrento, Italy: IEEE.

Lin, S.; Chen, H.; Wu, H.; Qiu, C.; and Lin, W. 2025. Temporal Query Network for Efficient Multivariate Time Series Forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*.

Lin, S.; Lin, W.; Hu, X.; Wu, W.; Mo, R.; and Zhong, H. 2024. CycleNet: Enhancing Time Series Forecasting through Modeling Periodic Patterns. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, 106315–106345. Vancouver, BC, Canada: Curran Associates, Inc.

Liu, H.; Ma, Z.; Yang, L.; Zhou, T.; Xia, R.; Wang, Y.; Wen, Q.; and Sun, L. 2023. SADI: A Self-Adaptive Decomposed Interpretable Framework for Electric Load Forecasting Under Extreme Events. In *Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. Rhodes Island, Greece.

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 *Proceedings of the 12th International Conference on Learning Representations*. Vienna, Austria.

Liu, Y.; Qin, G.; Huang, X.; Wang, J.; and Long, M. 2025a. Timer-XL: Long-Context Transformers for Unified Time Series Forecasting. In *Proceedings of the 13th International Conference on Learning Representations*. Singapore.

Liu, Z. 2025. FreqMoE: Enhancing Time Series Forecasting through Frequency Decomposition Mixture of Experts. In Li, Y.; Mandt, S.; Agrawal, S.; and Khan, E., eds., *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, 3430–3438. PMLR.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2025b. Kan: Kolmogorov-arnold networks. In *Proceedings of the 13th International Conference on Learning Representations*. Singapore.

Ma, X.; Li, X.; Fang, L.; Zhao, T.; and Zhang, C. 2024. U-mixer: An unet-mixer architecture with stationarity correction for time series forecasting. In *Proceedings of the 38th AAAI conference on Artificial Intelligence*, 13, 14255–14262.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Mallat, S. G. 2002. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7): 674–693.

Shi, X.; Wang, S.; Nie, Y.; Li, D.; Ye, Z.; Wen, Q.; and Jin, M. 2025. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. In *Proceedings of the 13th International Conference on Learning Representations*. Singapore.

Wang, S.; Li, J.; Shi, X.; Ye, Z.; Mo, B.; Lin, W.; Ju, S.; Chu, Z.; and Jin, M. 2025. Timemixer++: A general time series pattern machine for universal predictive analysis. In *Proceedings of the 13th International Conference on Learning Representations*. Singapore.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and ZHOU, J. 2024a. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *Proceedings of the 12th International Conference on Learning Representations*. Vienna, Austria.

Wang, Y.; Han, Y.; and Guo, Y. 2024. Self-adaptive Extreme Penalized Loss for Imbalanced Time Series Prediction. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 5135–5143. Jeju, South Korea.

Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Long, M.; and Wang, J. 2024b. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*.

Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; and Sun, L. 2023. Transformers in Time Series: A Survey. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 6778–6786. ijcai.org.

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 *Proceedings of the 11th International Conference on Learning Representations*. Kigali, Rwanda.

Xiu, Z.; Tao, C.; Gao, M.; Davis, C.; Goldstein, B. A.; and Henao, R. 2021. Variational disentanglement for rare event modeling. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 12, 10469–10477.

Zhang, G. P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50: 159–175.

Zhang, M.; Ding, D.; Pan, X.; and Yang, M. 2021. Enhancing time series predictors with generalized extreme value loss. *IEEE Transactions on Knowledge and Data Engineering*, 35(2): 1473–1487.