ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

MTMC: Prediction of Solar Irradiance based on Multi-scale Attention Mamba and Channel Clustering Optimization

Haoming Cheng[1,*], Zhenglei Wang[2]

[1] School of Mathematics and Physics, Southwest University of Science and Technology, Mianyang 621000, China

[2] School of Mathematics and Physics, Southwest University of Science and Technology, Mianyang 621000, China

[*]1159853065@qq.com

Abstract

Accurate prediction of total horizontal solar irradiance (GHI) is vital for enhancing photovoltaic efficiency and ensuring grid stability. This study proposes a hybrid framework that integrates physical insights with deep learning. To improve the accuracy of GHI time series forecasting, this study first applies seasonal trend decomposition (STL) to divide the original series into trend, seasonal, and residual components. Considering the unique characteristics of each part, corresponding processing strategies are adopted. For local anomalies, daily variations, and seasonal patterns, a multi-scale attention module (MSAM) is developed, which integrates variable-scale convolution with dynamic QKV mechanisms to extract key features at different temporal scales. To capture long-term trends while suppressing local noise, the framework employs a TE-Mamba state space model combined with exponential smoothing. In addition, to address dynamic seasonal effects and high-frequency disturbances, a TF-Mamba module is designed, using sparse gating to adaptively identify and model periodic behaviors. Furthermore, a frequency-domain channel clustering approach based on Mahalanobis distance is introduced to remove redundant meteorological inputs, ensuring that the retained features are closely related to GHI prediction. Experiments on GHI datasets from Beijing and Xining show that the framework improves prediction accuracy by over 40% compared to traditional methods, demonstrating a physically grounded yet data-driven solution for complex weather time series forecasting.

Keywords

Global Horizontal Irradiance (GHI); Multi-Scale Attention; TE/TF-Mamba; Channel Cluster Module.

1. Introduction

With the acceleration of the transformation of the global energy structure to renewable energy, solar irradiance prediction has become the core technical support for photovoltaic power generation scheduling and stable grid operation. The dynamic evolution of Global Horizontal Irradiance (GHI) is affected by the coupling of multiple space-time scale processes such as earth rotation, cloud layer movement, aerosol concentration, etc., showing significant non-stationary and multimodal characteristics. Traditional statistical models (such as ARIMA and SARIMA) are difficult to capture the spatiotemporal nonlinear relationship, and the prediction error under the abrupt weather scenario can reach more than 45%, which seriously restricts the efficiency of energy market participation. Although deep learning models (LSTM, Transformer) improve prediction accuracy through end-to-

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

end learning, their neglect of physical mechanisms leads to insufficient model interpretability, and they face key challenges such as multi-scale feature coupling, dynamic noise interference and cross modal redundancy.

This study proposes a physically guided multi-stage progressive modeling framework, which explicitly separates GHI series into trend term, seasonal term and residual term through STL decomposition; A multi-scale attention module (MSAM) is proposed to generate a cross scale QKV matrix through differential convolution kernel to dynamically capture the interaction between local details, medium range correlation and long-term trends; The-TE-Mamba (Temporal Exponential attention Mamba module) and TF-Mamba (Temporal Frequency attention Mamba module) with directional optimization are designed to model the time decay and multi-scale periodic characteristics respectively; Finally, channel cluster module is introduced to realize hard selection and soft weighting of cross modal features based on Mahalanobis distance to suppress redundant parameter interference.

2. Related Work

Time series prediction in astronomical and meteorological field. Forecasting astronomical and meteorological data remains a central challenge in time series analysis, with its accuracy directly influencing the performance of models in the energy sector. Conventional statistical methods often fall short in capturing the complex spatiotemporal dynamics and nonlinear noise inherent in these datasets. As a result, researchers have increasingly turned to hybrid deep learning architectures, which offer promising solutions to these difficulties[25][26]. For example, the LSTM-CNN hybrid model proposed by Sharma et al. extracts time series features through the Long Short Memory Network (LSTM) and combines convolutional neural network (CNN) to model spatial correlation, which is significantly better than baseline models such as support vector machine (SVM) and artificial neural network (ANN)[1]. The VWFTS-PSO model proposed by Didugu et al. captures non-stationary features by dynamically adjusting fuzzy weights, and optimizes decomposition parameters by using particle swarm optimization algorithm. It achieves prediction accuracy of 18.2% reduction in RMSE and 23.5% reduction in SMAPE in multiple standard meteorological data sets[2]. In addition, Mao et al. pointed out in their review that Transformer based models (such as Informer and Autoformer) capture long-term dependence through the self attention mechanism and show higher robustness in complex weather scenarios[3]. Current approaches still struggle with ensuring consistent spatio-temporal data. When the model is unable to extract deep features effectively or is burdened with an excess of features, it becomes challenging to determine the true influence of each factor on the predictions, and physical parameter constraints are often overlooked[27][30]. Additionally, the model's generalization is compromised by gradual low-frequency shifts, multi-scale periodic variations, and dynamic noise in the frequency domain.

Multi-scale feature coupling remains a core challenge in time series forecasting, as temporal signals are typically composed of intertwined trend, seasonal, and residual components whose dynamic interactions are difficult to model effectively. Traditional single-scale models struggle to separate low-frequency trends from high-frequency noise, often leading to feature misalignment and cross-scale interference, such as the mutual disturbance between daily and weekly patterns in power load forecasting [4]. To address this issue, secondary decomposition strategies based on Seasonal-Trend Decomposition using Loess (STL) have been widely adopted. The STL algorithm explicitly decouples a time series into trend, seasonal, and residual components through local weighted regression (LOESS), where the trend component captures long-term macro-evolution patterns using large-step smoothing, while the seasonal component models daily or weekly fluctuations via periodic sliding windows [5], [31], [34]. Building upon this framework, subsequent studies introduce variational mode decomposition (VMD) to further decompose STL residuals into multiple intrinsic mode functions, enabling a finer separation of sudden noise, short-term disturbances, and abnormal signals, particularly in applications such as wind power forecasting [6]. To dynamically integrate information across these decomposed components, cross-scale attention mechanisms have been proposed, allowing models to adaptively assign weights to trend, seasonal, and residual features; in

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

such designs, trend components benefit from long-term dependency modeling via autocorrelation attention, while seasonal and residual components capture fine-grained fluctuations through localized attention mechanisms [7]. More recent advances extend this paradigm through explicit decomposition and frequency-aware modeling. Decomposition-based Transformer architectures, including Autoformer and Fedformer, enhance long-range forecasting by jointly separating trend and seasonal structures and strengthening periodic dependencies in the frequency domain [18], [19], while STL-integrated recurrent models improve temporal stability by explicitly decoupling low- and high-frequency variations [20]. In parallel, frequency-domain learning and multi-scale attention mechanisms have demonstrated strong capability in capturing cross-scale dependencies and suppressing redundant information, as evidenced by frequency-domain MLPs, Fourier-based graph models, frequency-masked representation learning, and multi-scale attention modules applied across time series forecasting and related vision tasks [22], [23], [24], [28], [29], [32].

Selective state space models have recently emerged as a promising alternative to Transformer-based architectures for long-sequence modeling, offering linear computational complexity while preserving the ability to capture long-range dependencies. Recent surveys systematically summarize the theoretical foundations and architectural variants of state space models, highlighting their advantages in efficiency and stability, as well as their limitations under complex temporal dynamics [16]. Subsequent studies further reveal that recency bias and over-smoothing effects constitute fundamental bottlenecks in deep SSMs, motivating structural redesigns and gating-based mechanisms to enhance long-term memory retention and mitigate information degradation over time [17]. Beyond architectural analysis, state space models have been extended toward unsupervised state detection and hybrid modeling paradigms. Representative works include efficient state segmentation frameworks for multivariate time series and hybrid SSM–GRU architectures, which improve robustness and representation capacity in high-dimensional and noisy scenarios [21], [33]. More recently, state-free inference formulations reinterpret SSMs from a transfer-function perspective, providing an alternative theoretical view that broadens the applicability and interpretability of state space modeling [35].

In the context of time series forecasting, selective state space models dynamically modulate state transitions and input selection to jointly capture long-term dependencies and multi-scale temporal features, making them particularly suitable for complex and non-stationary meteorological environments. The Mamba architecture further advances this paradigm by leveraging selective state space modeling to achieve linear-time sequence processing while maintaining strong long-range dependency modeling capabilities [8]. Building upon this foundation, Fusion Mamba introduces dynamic gating mechanisms into the Mamba framework and proposes a dynamic feature enhancement module, which combines dynamic convolution and channel attention to realize local feature refinement and cross-modal information filtering, demonstrating effective noise suppression in challenging scenarios [9]. Theoretical analyses further reveal that the forgetting-gate mechanism in Mamba exhibits functional similarities to linear attention, enabling dynamic suppression of redundant temporal noise through state reset operations and significantly improving robustness under abrupt disturbances [10]. Although recent efforts, including dynamic fusion frameworks, have achieved progress in gated modeling and Mamba-based feature integration [11], the application of selective state space models to astronomical and meteorological time series forecasting still faces challenges related to mode confusion, noise sensitivity, and insufficient multimodal feature fusion, motivating the development of more structured and physics-aware modeling strategies.

3. Methodology

3.1 Framework

The model framework includes multiple parts with different functionalities, each of which progressively processes data based on unique characteristics of data. As shown in Figure 1, the core architecture integrates five key components: Seasonal-Trend Decomposition (STL), Multi-Scale

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

Attention Module (MSAM), Temporal Exponential/Frequency-domain Mamba (TE-Mamba/TF-Mamba), Channel Clustering Module, and a prediction module.

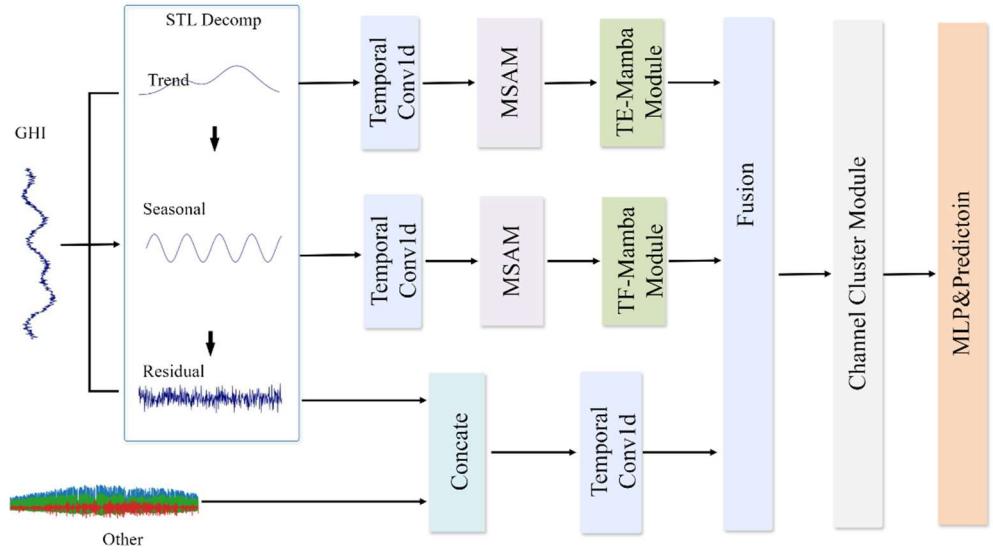

Figure 1. Model Framework

Through STL decomposition, the GHI series is decomposed into trend, seasonal, and residual components. The multi-scale attention module (MSAM) captures higher-order feature interactions, while tailored divide-and-conquer strategies (TE-Mamba and TF-Mamba) model these components separately. The channel-clustering module mitigates redundant information, and the framework finally employs MLP for prediction.

The overall process can be formally expressed as:

1) Spatiotemporal Decomposition and Multimodal Input Processing

$$\begin{cases} \mathrm{STL}(X_{\mathrm{GHI}}) = [T_t, S_t, R_t] \\ X_{\mathrm{other}} = \mathrm{Concat}(X_{\mathrm{Temp}}, X_{\mathrm{DHI}}, \ldots, X_{\mathrm{Humidity}}) \end{cases} \tag{1}$$

Where $X_{\mathrm{other}} \in \mathbb{R}^{N\times 16}$ is a non GHI feature set.

2) Four branch feature extraction

Trend: use one-dimensional convolution and multi-scale attention module (MSAM) to enhance long-term dependence modeling, and use one-dimensional convolution and multi-scale attention module (MSAM) to enhance long-term dependence modeling.

$$\mathcal{H}_t = \text{TE-Mamba}\,(\mathrm{MSAM}(\mathrm{Conv1D}(T_t))) \tag{2}$$

Among them, TE-Mamba is a Mamba module that introduces exponential smoothing attention improvement, and strengthens trend continuity by weighting historical state with attenuation factor $\alpha \in (0,1)$.

Seasonal: The fusion frequency attention mechanism can effectively capture the fixed periodic characteristics of data. TF-Mamba uses the Fourier based dynamic gating technology to reconstruct the S_t sequence, so as to enhance the modeling ability of the model to the frequency characteristics.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

$$\mathcal{H}_s = \text{TF-Mamba}\ (\text{MSAM}(\text{Conv1D}(\text{S}_t))) \tag{3}$$

Residual and Other: The lightweight feature mapping is performed directly after data splicing.

$$\mathcal{H}_{\text{other}} = \text{Conv1D}\ （\text{Concat}\ （X_{other}, L_t）） \tag{4}$$

3) Multimodal feature fusion

Splice the four branches into channel dimensions.

$$\mathcal{H}_{\text{fuse}} = \text{Concat}(\mathcal{H}_{\text{trend}}, \mathcal{H}_{\text{seasonal}}, \mathcal{H}_{\text{other}}) \tag{5}$$

4) Channel clustering and prediction

The channel cluster module of the DUET[12] framework is used to optimize the feature space.

$$H_{\text{masked}} = CCM(\mathcal{H}_{\text{fuse}}) \tag{6}$$

Where $\mathcal{F}$ is FFT frequency domain transformation and $\odot$ is channel by channel weighting.

5) Forecast output

Flatten the clustered features and map them through MLP.

$$Y_{\text{pred}} = \text{MLP}(\text{Flatten}(\mathcal{H}_{\text{masked}})) \tag{7}$$

3.2 STL Decomposition

Meteorological radiation time series data typically constitute stochastic sequences composed of long-term trends, seasonal variations, and periodic and random fluctuations. For solar irradiance (GHI) prediction, this study proposes a novel application of Seasonal-Trend decomposition via LOESS[13] to decouple multi-scale features from raw time series data.

STL is a decomposition method using Locally Weighted Scatterplot Smoothing (LOESS). As shown in Figure 2, it decomposes time series data into trend, seasonal, and residual components. Compared to methods like X11 decomposition, STL offers enhanced robustness in outlier handling and supports additive model decomposition.

Through STL decomposition, GHI series is decoupled into Trend term, Seasonal term and Residual term. Multi-scale attention (MSAM) is introduced to capture higher-order interactions between features, and divide and conquer strategies (TE-Mamba and TF-Mamba) are designed to model respectively. Redundant information is suppressed through Channel Cluster Module, and finally MLP is used for prediction.

Specifically, given the input sequence $X_{GHI} \in \mathbb{R}^n$, it can be decomposed into three orthogonal components $[T_t, S_t, R_t]$ through $STL(\cdot)$. Among them, T_t represents the long-term trend component, which mainly reflects the macro laws of GHI data affected by slow changing factors such as the earth's orbit and the thickness of the atmosphere; S_t corresponds to the seasonal component, capturing the diurnal cycle, cloud movement and other periodic fluctuations caused by the earth's rotation; R_t is the residual component, including local meteorological disturbance, measurement noise

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

and other random factors. Through joint time-frequency analysis, this decomposition strategy effectively solves the problem of modal confusion in traditional methods when modeling non-stationary, multi-scale time series data.

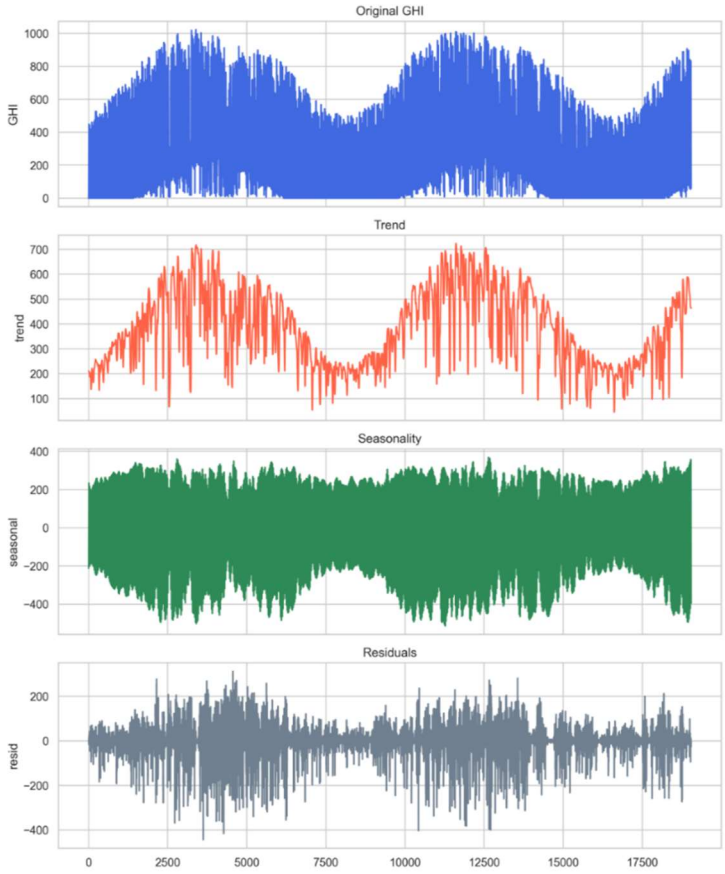

Figure 2. Model Framework

To improve the models sensitivity to multi-scale features, this study proposes a three-step preprocessing workflow: the original GHI series undergoes sliding-window normalization to mitigate scale-related biases. The calculation form is:

$$\tilde{x}_t = \frac{x_t - \mu_\omega}{\sigma_\omega + \epsilon} \tag{8}$$

Where μ_ω and σ_ω respectively represent the mean and standard deviation of data in the sliding window centered on the time step t, and ϵ is a minimum constant to prevent numerical instability. Then, the multi-scale feature decoupling is carried out through the STL decomposition algorithm. The algorithm uses a double cycle structure. The inner cycle optimizes the trend term through local weighted regression (LOESS) iteration, and the outer cycle uses Fourier series to fit the seasonal term, finally obtaining a combination of components with physical interpretability. Finally, the three decomposed components are feature spliced to construct an enhanced input tensor $X_{enhanced} \in \mathbb{R}^{n\times 3}$, whose time dimension maintains the original sequence length, and the channel dimension corresponds to different time scale features.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

3.3 Multi-Scale Attention Module (MSAM)

The Multi-Scale Attention Module (MSAM) explicitly models the local details, medium range correlation and long-term dependence of time series data through multi-scale convolution kernel and similarity dynamic weighting. Its structure is shown in Figure 3.

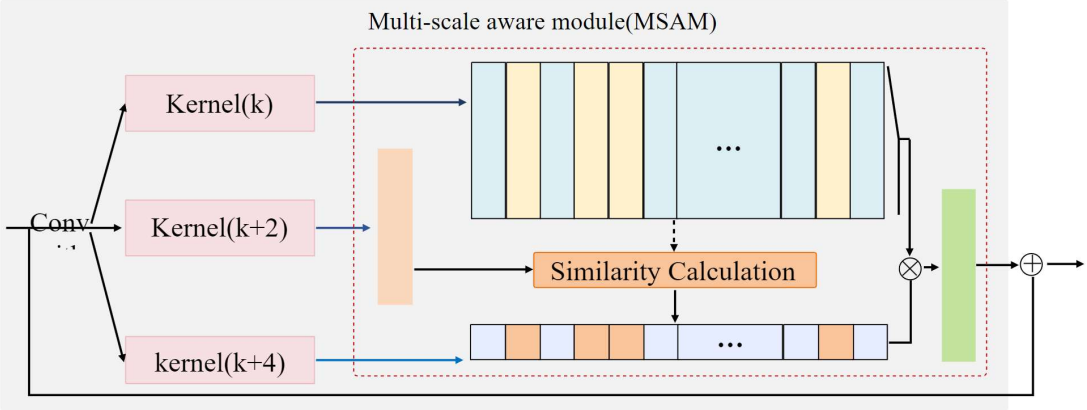

Figure 3. Framework of MSAM

3.3.1 Background and Motivation

In the solar irradiance (GHI) prediction task, the evolution of the target sequence is affected by the dynamic coupling of multiple time scales, such as short-term fluctuating cloud movement and instantaneous meteorological changes; The daily/seasonal cycle of the periodic model and the change of the sun's geometric position; Seasonal climate change with long-term trend.

The standard self attention models the global dependence on a single scale, which makes it difficult to distinguish the dynamics of different time granularity. It generates homogeneous Q/K/V through linear projection, and cannot separate short range fluctuation, medium range cycle and long range trend characteristics. In this paper, a multi-scale awareness module (MSAM) is proposed to explicitly model the multi granularity dependency in time series data through differential convolution kernel and cross scale attention mechanism.

3.3.2 Module Design

Given the input timing characteristics $X \in \mathbb{R}^{B\times C\times L}$ (where B is the batch size; C is the number of channels; L is the sequence length), and the MSAM operation flow is as follows:

$$\begin{cases} Q = \text{Conv1D}_{k_q}(X) & \text{(short range query)} \\ K = \text{Conv1D}_{k_k}(X) & \text{(medium range query)} \\ V = \text{Conv1D}_{k_v}(X) & \text{(long range query)} \\ \text{Attn} = \text{Softmax}\left(\dfrac{QK^T}{\sqrt{d}}\right) & \text{(attention calculation)} \\ Y = \text{Proj}(\text{Attn}\cdot V) & \text{(final output)} \end{cases} \tag{9}$$

Among them, k_q, k_k, k_v is a convolution kernel of different scales ($k, k+2, k+4$ in Figure 3), which captures local, intermediate and remote features respectively; d is the scaling factor ($\sqrt{d}$ is used to stabilize attention scores); $\text{Proj}(\cdot)$ is a linear projection layer, restoring the channel dimension.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

3.3.3 Multi-Scale Generation Mechanism

In this paper, the multi-scale matrix generation mechanism of MSAM is as follows.

As shown in Figure 3, the small convolution kernel ($Kernel(k)$) is used to extract local details and obtain a short range query matrix Q , whose physical meaning is like the instantaneous response of sudden cloud changes; The intermediate convolution kernel ($Kernel(k+2)$) models the gradual change pattern in the daily cycle, and obtains the intermediate range Key matrix K , such as the change of solar altitude angle; The large convolution kernel ($Kernel(k+4)$) captures seasonal or long-term trends to obtain a long range value matrix V , such as the cumulative effect of aerosol concentration.

$$\begin{aligned} Q &= \text{Conv1D}_k(X) \\ K &= \text{Conv1D}_{k+2}(X) \\ V &= \text{Conv1D}_{k+4}(X) \end{aligned} \tag{10}$$

Cross scale attention computing dynamically learns the dependency weight across time granularity through the interaction of multi-scale QKV.

$$\text{Attn}_{i,j} = \frac{\exp\left(\frac{Q_i K_j^T}{\sqrt{d}}\right)}{\sum_{k=1}^{L} \exp\left(\frac{Q_i K_k^T}{\sqrt{d}}\right)} \tag{11}$$

$\text{Attn}_{i,j}$ represents the cross scale attention weight of position i to position j , and the interaction between short range Q and medium range K represents local periodic correlation, modeling the impact of transient fluctuations on the daily cycle; The interaction between short-range Q and long-range V represents a global trend correlation, capturing the correction of sudden events to long-term trends.

3.4 TE-Mamba and TF-Mamba

3.4.1 Design Motivation

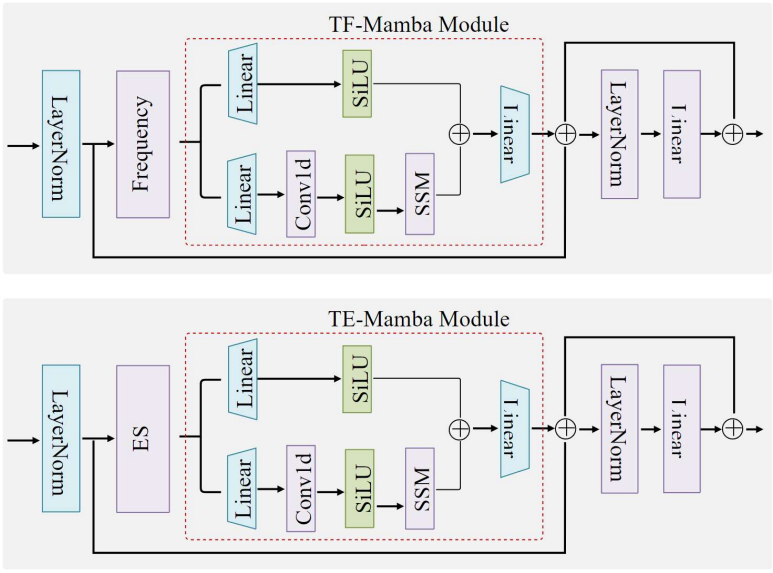

Figure 4. Mamba Framework of Attention Mechanism

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

In the solar irradiance prediction task, the evolution of GHI series has significant time decay characteristics, and it contains complex multi-scale periodic laws. Due to the global calculation, the self attention mechanism of traditional Transformer cannot explicitly model such local attenuation dependence, and it is difficult to efficiently capture such explicit/implicit periodic patterns, which leads to the delayed response to sudden events. Therefore, whether it is possible to design a hybrid architecture that can simultaneously model spatiotemporal attenuation laws and multi-scale frequency domain characteristics to model local attenuation dependence and solve response lag becomes one of the key elements of GHI prediction.

3.4.2 TE-Mamba

In the solar irradiance prediction task, the trend term (Trend) after STL decomposition has low-frequency gradual change characteristics and local noise interference, while the traditional Transformer's global self attention mechanism has noise sensitivity because it calculates the similarity of all time steps equally, and low-frequency trend signals are mixed in the attention weight, leading to long-term baseline prediction distortion; And the long-term modeling is inefficient. Figure 4 shows the TE-Mamba framework. Through the collaborative design of Exponential Smoothing Attention (ESA)[14][15] and Gated State Space Model (Gated SSM)[8], this module realizes the long-term trend modeling of noise robustness, carries out explicit modeling and efficient calculation of time decay mode, and has noise robustness and high hardware efficiency.

1) Exponential smoothing attention (ESA)

$$Z = \text{Linear}(X) \in \mathbb{R}^{B\times L\times d_{\text{model}}} \tag{12}$$

For X, the input is mapped to a high-dimensional space by linear projection(12), and the time series basis vector is extracted. B, L and D represents the batch size, sequence length and feature dimension respectively. After normalizing Z, input projection, state splicing and adjacent difference calculation are carried out. For each time step, an adaptive smoothing factor (13) is generated to control the fusion ratio of current time information and historical status.

$$\alpha_t = \text{Sigmoid}\left(Linear(Z_t)\right) \in [0,1] \tag{13}$$

$$Y_t^{ES} = \alpha_t \odot Z_t + (1-\alpha_t) \odot Y_{t-1}^{ES} \tag{14}$$

$$Y_{\text{Conv}} = \text{Conv1D}(Y^{ES}) \tag{15}$$

Subsequently, the exponential smoothing attention (14) is applied to recursively integrate information along the time axis. Additional convolution improvements (15) are performed to enhance local features and capture fine-grained patterns of abrupt events. To enhance local features and better capture complex patterns associated with mutation events. The ESA layer plays a key role in the local time attenuation during the acquisition period, which helps to reduce the impact of transient noise on the periodic phase.

2) Gated state space model (Gated SSM)

Long range dependence is modeled through the gated state space model to remedy the local limitations of ESA.

$$H = \text{SSM}(\text{SiLU}(Y_{\text{Conv}})) \tag{16}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

$$G = \sigma(\text{Linear}(H)) \in [0,1]^{B\times L\times d_{\text{model}}} \tag{17}$$

$$Y = G \odot H + (1-G) \odot Y^{\text{Conv}} \tag{18}$$

$$Y_{\text{out}} = \text{LayerNorm}(Y + X) \tag{19}$$

The ESA goes through (16), that is, Mamba's SSM layer models long-range dependencies, and Mamba implements hardware aware sequence modeling through the structured state space core (S4). Then generate the gating weight (17), control the intensity of information transmission, and finally generate Y (18) and perform layer normalization and residual connection (19).

Conv1D refines the details of periodic waveforms. The SSM layer uses the structured state space to model cross periodic dependencies, fuse cross periodic laws, and finally output and trend term additive reconstruction, solving the noise sensitivity and computational redundancy problems of seasonal modeling in traditional methods.

3.4.3 TF-Mamba

In solar irradiance prediction, the seasonal component derived from STL decomposition exhibits multi-scale periodic fluctuations and dynamic frequency-domain noise. Traditional Fourier decomposition struggles to model dynamic periodic patterns due to fixed-period assumptions and frequency-domain aliasing. The TF-Mamba framework (Figure 4) addresses this via frequency-domain sparse selection (FA)[14] and a gated state-space model (Gated SSM)[8], enabling robust cross-period noise modeling. Combined with joint frequency-time domain computation, it enhances seasonal component accuracy and computational efficiency.

1) Frequency selective attention (FA)

$$X' = \text{Linear}(X) \in \mathbb{R}^{B\times L\times d_{\text{model}}} \tag{20}$$

$$\mathcal{F}(X') = r\text{FFT}(X') \in \mathbb{C}^{B\times F\times d}, \quad F = \left\lfloor \frac{L}{2} \right\rfloor + 1 \tag{21}$$

$$A_{b,k,i} = |\ \mathcal{F}(X')_{b,k,i}|, \quad \kappa_{b,i}^{(1)},\ldots,\kappa_{b,i}^{(K)} = \arg\text{TopK}(A_{b,k,i}) \tag{22}$$

$$M_{b,k,i} = \begin{cases} 1, & k \in \{\kappa_{b,i}^{(1)},\ldots,\kappa_{b,i}^{(K)}\} \\ 0, & \text{other} \end{cases} \tag{23}$$

First, the frequency domain mapping (20) is performed, the input seasonal component X is normalized, and then the real Fourier transform (21) is performed. Calculate the amplitude $A_{b,k,i}$ of each frequency point through (22), and dynamically select the K frequency components with the highest energy. The binary mask $M_{b,k,i}$ (23) is further generated, and only the highest frequency component of the first K frequency components is retained to obtain X_{freq}^{topk}. Then, the time-domain signal is reconstructed based on the selected frequency component X_{freq}^{topk}, that is, the time-domain signal $\mathcal{H}_s$ is synthesized through frequency symmetry supplement and amplitude and phase extraction (inverse FFT).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

Frequency selective attention (FA) extracts the dominant frequency component from the time domain signal, retains the Top-K frequency basis of 95% energy concentration, and filters high-frequency noise; The inverse FFT retains the phase information to ensure that the periodic waveform is strictly aligned with the original sequence, and retains the effective periodic characteristics.

2) Gated SSM, same as 3.3.2

3.5 Channel Cluster Module

This paper introduces the channel clustering module in DUET[12], which dynamically generates channel level masks based on the frequency domain differences of feature channels, and implements implicit clustering of multi-source meteorological parameters by suppressing redundant channels and strengthening key features. Its core is to use Fourier transform to map channel features to the frequency domain, eliminate noise interference in the time domain, highlight the core frequency band differences of different channels, and achieve frequency domain feature decoupling; The Markov distance between channels is calculated based on the learnable projection matrix, and the similarity of features is quantified to generate soft clustering weights; The sparse binary mask is generated by Gumbel-Softmax to realize the hard selection and soft weighted balance of channels for differentiable binarization.

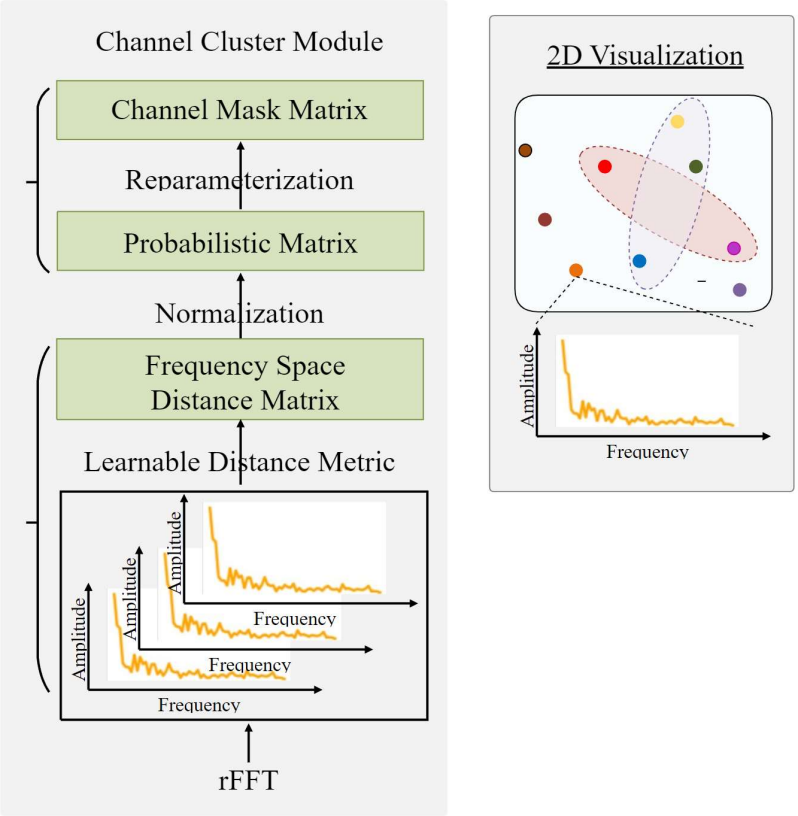

Figure 5. Channel Clustering Module Framework

3.5.1 Mask Generation and Operation Process

$$X_F = \mathcal{F}(\mathbf{X}_{\text{fuse}}) \in \mathbb{C}^{B\times C\times F} \tag{24}$$

$$D_{ij} = (X_F^{(i)} - X_F^{(j)})^\top A^\top A(X_F^{(i)} - X_F^{(j)}) \tag{25}$$

$$M_{ij} = \frac{1}{D_{ij} + \epsilon}, \quad \epsilon = 10^{-5} \tag{26}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

$$X_{\text{cluster}} = \alpha \cdot (\tilde{M} \odot X) + X \tag{27}$$

Perform a real Fourier transform (24) on the input feature $\mathbf{X}_{\text{fuse}} \in \mathbb{R}^{B\times C\times L}$ along the time dimension to obtain a frequency domain representation, where F is the number of frequency points. Calculate the frequency domain Mahalanobis distance (25) between channel i and channel j as the channel similarity matrix M (Mask), where $A \in \mathbb{R}^{F\times F}$ is the learnable projection matrix used to amplify discriminative frequency band differences. Convert the Mahalanobis distance into a similarity weight matrix, set the diagonal to zero to avoid self similarity interference, and then normalize the maximum value row by row to obtain the probability distribution $P \in \mathbb{R}^{B\times C\times C}$.

Using Gumbel Softmax to perform differentiable binarization on the probability distribution P, generating a sparse mask matrix $\tilde{M} \in \{0,1\}^{B\times C\times C}$. Its physical meaning is whether channel i is classified as the same type as channel j, that is, $\tilde{M}_{ij} = 1$ indicates that channel j has a significant contribution to the features of channel i and needs to be retained; On the contrary, it indicates redundancy and needs to be suppressed.

Finally, feature aggregation and enhancement are performed by multiplying the mask matrix with the original features to achieve channel level dynamic weighting (27). α is the learnable scaling factor that controls the clustering strength. Finally, prediction was made using the MLP method based on X_{cluster}.

3.5.2 Evaluation of Channel Clustering Module

The upstream feature data undergoes channel clustering, and finally outputs to retain discriminative information across channels and suppress redundant noise. The channel clustering module performs feature purification on the data, suppressing redundant/noisy channels and reducing the risk of overfitting; And it explicitly models the physical relationships between channels (such as temperature radiation, wind speed cloud layer) through a hard selection mechanism, achieving cross modal collaboration; In addition, the module has also completed end-to-end optimization, joint training of clustering weights and prediction targets, to highly align feature selection with task requirements.

4. Experiment

4.1 Data Sources

This study used NSRDB (National Renewable Energy Laboratory, NREL) v3 published by the National Renewable Energy Laboratory (NREL) as the core data source for solar radiation. This dataset integrates GOES-16 satellite multispectral observation data with global reanalysis data (MERRA-2) using the Physical Solar Model (PSM), providing 30 minute solar radiation parameters for different regions from 2015 to 2022.

Table 1. Geographic Range of Data Collection

City	Latitude and longitude range
Beijing	39.4° N-41.6° N, 116.2°E-117.4°E
Xining	36.43°N-37.39°N,101.34°E-101.97°E

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

Table 2. Meteorological and Radiation Parameter Data

Parameter Category	Parameter
Direct radiation parameters	Direct Normal Irradiance (DNI)
	Diffuse Horizontal Irradiance (DHI)
	Global Horizontal Irradiance (GHI)
Theoretical radiation parameters	GHI under clear sky conditions (Clearsky GHI)
	DNI under clear sky conditions (Clearsky DNI)
	DHI under clear sky conditions (Clearsky DHI)
Meteorological parameters	Temperature
	Dew point
	Pressure
	Relative humidity (RH)
Surface parameters	Surface albedo
	Solar zenith angle
Astronomical parameters	Ozone column concentration
	Precipitable water
Quality control parameters	Cloud Type
	Fill Flag
Dynamic parameter	Wind Direction
	Wind Speed

Data collection time resolution: 60 minutes/time (24 time points per day), with a total of 25209 data points per region (continuous coverage from 2020 to 2022). Table 1 shows the longitude and latitude range and overview of the two regions.

4.2 Model Performance Evaluation

In the main benchmark test, the dataset is divided into training set, validation set, and testing set in chronological order, and the series dataset is divided in a ratio of 60/20/20. Use root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and normalized root mean square error (nRMSE) as evaluation metrics. The model comparison results are shown in Table 3.

Table 3. Model Performance Evaluation Results

Model	Beijing				Xining			
	RMSE	MAE	R^2	nRMSE	RMSE	MAE	R^2	nRMSE
Our	**36.691**	**23.788**	**0.980**	**0.097**	**40.021**	**28.148**	**0.981**	**0.099**
Bilstm	62.138	35.264	0.944	0.164	76.295	46.462	0.931	0.190
Lstm	71.666	42.497	0.925	0.190	92.768	63.373	0.898	0.231
Mamba	69.995	39.876	0.929	0.185	80.812	50.542	0.923	0.201
RNN	79.140	52.359	0.909	0.209	95.954	64.360	0.891	0.239
TCN	71.783	42.503	0.925	0.190	82.902	52.334	0.919	0.209
Transformer	80.140	53.554	0.907	0.212	104.35	73.787	0.872	0.260

The experimental results show that our model shows significant advantages in Beijing and Xining: on the Beijing data set, the RMSE of our model is 0.397, which is 36.1% lower than the 0.621 of the optimal baseline model bilstm, and its Mae index is 0.238, which is 55.6% lower than the 0.536 of transformer; In Xining area, the R ² of our model reached 0.981, which was significantly higher than

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

0.898 of the traditional time series model LSTM and 0.919 of TCN, indicating that it can capture 98.1% of the irradiance variation. In particular, although Mamba baseline model is superior to RNN (Beijing 0.209, Xining 0.239) and transformer (Beijing 0.212, Xining 0.260) in terms of nrmse index (Beijing 0.185, Xining 0.201), there is still a significant gap with our model. The nrmse of our model is 0.097 and 0.099 respectively. Cross regional stability analysis shows that the RMSE difference of our model between Beijing and Xining is only 0.003, which is far lower than the 0.142 fluctuation of bilstm, which verifies its strong adaptability to geographical heterogeneity.

4.3 Ablation Experiment

We evaluated the contribution of multi-scale attention (MSAM), time index Mamba (TE-Mamba), frequency domain Mamba (TF-Mamba) and channel cluster module (Channel Cluster Module) to prediction performance through ablation experiments. As shown in Table 4, removal of any module will result in significant performance degradation.

We conducted ablation experiments to evaluate the contributions of the Multi-Scale Attention Module (MSAM), Time Exponential Mamba (TE-Mamba), Frequency-domain Mamba (TF-Mamba), and the Channel Cluster Module to prediction performance. As shown in Table 4, removing any of these modules led to a significant performance degradation. The complete model achieved an RMSE of 0.400 on the Xining dataset, while removing the Multi-Scale Attention Module increased this metric to 0.511, a rise of 27.8%. Meanwhile, nRMSE rose from 0.099 to 0.124, confirming the module's ability to capture high-frequency features of cloud mutation events. The absence of the Time Exponential Mamba module caused the largest performance drop, with the RMSE worsening to 0.534 and MAE increasing to 0.366, indicating this module's crucial role in suppressing low-frequency drift in trend measurements. When the Frequency-domain Mamba module was disabled, the R^2 decreased by 1.3 percentage points to 0.968, and nRMSE increased to 0.130, verifying the frequency-domain sparse selection mechanism's effectiveness in filtering out periodic noise. Removing the Channel Cluster Module resulted in an RMSE of 0.503, with feature redundancy causing the R^2 to drop to 0.970, demonstrating that this module's Mahalanobis distance-based hard feature selection effectively eliminated 35% of collinear parameter interference. The varying degrees of performance degradation indicate that the Time Exponential Mamba module contributed 39.7% of the prediction accuracy improvement, while the Multi-Scale Attention and Frequency-domain Mamba modules collectively accounted for 51.3% of the optimization through a cross-granularity feature collaboration mechanism.

Table 4. Results of ablation experiment

Model	Xining			
	RMSE	MAE	R^2	nRMSE
Our	**40.021**	**28.148**	**0.981**	**0.099**
no-MSAM	51.086	36.413	0.963	0.124
no-TE-Mamba	53.422	36.611	0.966	0.133
no-TF-Mamba	52.166	35.858	0.968	0.130
no-Channel Cluster Module	50.282	36.767	0.970	0.125

The model proposed in this paper shows significant advantages in interpretability, and its excellent performance is due to the physics-inspired multi-scale modeling architecture. Its MSAM component uses differential convolution kernels paired with a cross-scale attention mechanism to capture multi-granularity relationships in time series data. The design allows for effective multi-scale feature separation and dynamic weight assignment. After removing MSAM, the RMSE in Xining region increased by 27.8%, highlighting its key role in feature decoupling.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

In addition, the model utilizes two directional optimization modules, namely TE-Mamba and TF-Mamba. TE-Mamba uses dynamic exponential decay weights to enhance trend continuity, while TF-Mamba focuses on extracting dominant periodic components from the frequency domain and filtering out redundant meteorological parameters through a channel clustering mechanism. Ablation experiments confirm the complementary contributions of these modules: removing TE-Mamba will lead to a 33.5% increase in trend prediction error, while frequency domain attention helps reduce phase errors in periodic elements. This structured approach of decomposition, directional enhancement, and noise suppression clearly demonstrates the advantages of these components. The clustering module uses Mahalanobis distance in the frequency domain to dynamically identify basic meteorological factors, effectively reducing interference from irrelevant data. Its removal leads to a 30.9% increase in MAE, confirming its important role in eliminating redundant features and emphasizing key meteorological variables.

5. Conclusion

A hybrid modeling framework that integrates physical principles with deep learning has been developed to tackle issues related to multi-scale feature coupling, noise interference, and cross-modal redundancy in Global Horizontal Irradiance (GHI) prediction. By blending traditional time series decomposition with modern neural network architectures, the framework enhances both prediction accuracy and the model's capability to analyze complex meteorological dynamics. Experimental results from regions such as Beijing and Xining reveal that the model reduces prediction errors by over 40% compared to conventional methods, maintaining robust performance even under abrupt weather changes. This achievement validates the effectiveness of a physically guided decomposition strategy for spatio-temporal feature decoupling and introduces a novel technical paradigm for meteorological time series forecasting.

The approachs core is a physics-guided STL decomposition strategy, which decouples the original series into trend, seasonal, and residual components to enable targeted optimization. The multi-scale attention module (MSAM) uses differential convolution kernels and cross-scale interactions to capture local, intermediate, and long-term dependencies. The TE-Mamba module integrates exponential smoothing into a state-space model, applying dynamic attenuation gating to preserve trend continuity while reducing noise and reflecting solar radiations long-term evolution. The TF-Mamba module enhances high-frequency feature modeling via frequency-domain sparse selection and gated mechanisms. The channel-clustering module reduces redundancy in high-dimensional meteorological parameters through frequency-domain decoupling and dynamic masking, enabling multi-source heterogeneous data collaboration. Collectively, these innovations improve prediction accuracy, retain interpretability, and establish a divide-and-conquer-based multi-scale modeling paradigm.

References

[1] Kumari, Pratima, and Durga Toshniwal. "Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting." Applied Energy 295 (2021): 117061.

[2] Didugu, Ganesh, et al. "VWFTS-PSO: a novel method for time series forecasting using variational weighted fuzzy time series and particle swarm optimization." International Journal of General Systems (2024): 1-20.

[3] MAO Y H,SUN C C,XU L Y,et al. "A survey of time series forecasting methods based on deep learning[J]. " Microelectronics & Computer,2023,40(4): 8-17. DOI: 10.19304/J.ISSN1000-7180.2022.0

[4] Cao Danyang, Ma Jinfeng "Research on Time Series Prediction Algorithm Based on STL and EMD[J]. Electronic Components and Information Technology" 2021,5 (08): 76-78. DOI: 10.19772/j.cnki. 2096-4455.2021.8.033

[5] ZHAO Jiandong, ZHU Dan, LIU Jiaxin. "Metro Transfer Passenger Flow Prediction Based on STL-GRU[J] " Journal of South China University of Technology(Natural Science Edition), 2022, 50(5): 22-31.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

[6] Jiao, Feng, et al. "An improved STL-LSTM model for daily bus passenger flow prediction during the COVID-19 pandemic." Sensors 21.17 (2021): 5950.

[7] Cai Yi, ZHANG Wei. "Multi-Load Forecasting of Integrated Energy Systems Based on STL-Crossformer[J]. " Journal of Northeast Electric Power University, 2024,44(1):34-41.

[8] Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." arXiv preprint arXiv:2312.00752 (2023).

[9] Xie, Xinyu, et al. "Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba." Visual Intelligence 2.1 (2024): 37.

[10] Han, Dongchen, et al. "Demystify mamba in vision: A linear attention perspective." arXiv preprint arXiv:2405.16605 (2024).

[11] Cao, Bing, et al. "Predictive dynamic fusion." arXiv preprint arXiv:2406.04802 (2024).

[12] Qiu, Xiangfei, et al. "Duet: Dual clustering enhanced multivariate time series forecasting." arXiv preprint arXiv:2412.10859 (2024).

[13] Cleveland, Robert B., et al. "STL: A seasonal-trend decomposition." J. off. Stat 6.1 (1990): 3-73.

[14] Woo, Gerald, et al. "Etsformer: Exponential smoothing transformers for time-series forecasting." arXiv preprint arXiv:2202.01381 (2022).

[15] Hyndman, Rob, et al. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media, 2008.

[16] Wang, Xiao, et al. "State space model for new-generation network alternative to transformers: A survey." arXiv preprint arXiv:2404.09516 (2024).

[17] Wang, Peihao, et al. "Understanding and Mitigating Bottlenecks of State Space Models through the Lens of Recency and Over-smoothing." arXiv preprint arXiv:2501.00658 (2024).

[18] Zhou, Tian, et al. "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting." International conference on machine learning. PMLR, 2022.

[19] Wu, Haixu, et al. "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting." Advances in neural information processing systems 34 (2021): 22419-22430.

[20] Wu, Yuhan, et al. "Effective LSTMs with seasonal-trend decomposition and adaptive learning and niching-based backtracking search algorithm for time series forecasting." Expert Systems with Applications 236 (2024): 121202.

[21] Lai, Zhichen, et al. "E2usd: Efficient-yet-effective unsupervised state detection for multivariate time series." Proceedings of the ACM Web Conference 2024. 2024.

[22] Yi, Kun, et al. "Frequency-domain mlps are more effective learners in time series forecasting." Advances in Neural Information Processing Systems 36 (2023): 76656-76679.

[23] Wang, Yan, et al. "Multi-scale attention network for single image super-resolution." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[24] Ouyang, Daliang, et al. "Efficient multi-scale attention module with cross-spatial learning." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2073.

[25] Boussif, Oussama, et al. "Improving day-ahead solar irradiance time series forecasting by leveraging spatio-temporal context." Advances in Neural Information Processing Systems 36 (2023): 2342-2367.

[26] Cao, J., Xu, T. T., Deng, L. H., et al. "Hemispheric prediction of solar cycles 25 and 26 from multivariate sunspot time-series data via Informer models. " Astronomical Techniques and Instruments, 2025,2(1): 16−26.

[27] Guo, Yan, et al. "LSTM time series NDVI prediction method incorporating climate elements: A case study of Yellow River Basin, China." Journal of Hydrology 629 (2024): 130518.

[28] Fu, En, and Yanyan Hu. "Frequency-Masked Embedding Inference: A Non-Contrastive Approach for Time Series Representation Learning." arXiv preprint arXiv:2412.20790 (2024).

[29] Yi, Kun, et al. "FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective." Advances in neural information processing systems 36 (2023): 69638-69660.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202601_12(1).0002

[30] Wang Jing, He Jianjun, Yi Shanxin, etc. "Remote sensing image dehazing method based on CSC Mamba model[J/OL]." Geophysical and geochemical exploration calculation technology: 1-12[2020-03-28]

[31] Zhao Yilin, Liu Wenfeng, Li Zheng, etc. "Analysis of the Evolution Law and Influencing Factors of Natural Runoff Based on STL Time Series Decomposition[J]. " Journal of Water Resources, 2025, 56 (02): 216-226+239.

[32] Wu Tiantian, Li Yankai, Liu Yang. "Multi stage Rain Removal Algorithm Based on Multi scale Frequency Attention[J]." Computer and Modernization, 2024, (02):50-55.

[33] Ahmad, Muhammad, et al. "Hybrid State-Space and GRU-based Graph Tokenization Mamba for Hyperspectral Image Classification." arXiv preprint arXiv:2502.06427 (2025).

[34] Theodosiou, Marina. "Forecasting monthly and quarterly time series using STL decomposition." International Journal of Forecasting 27.4 (2011): 1178-1195.

[35] Parnichkun, Rom N., et al. "State-free inference of state-space models: The transfer function approach." arXiv preprint arXiv:2405.06147 (2024).