Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

TimeMixer-ME: A Dual-Memory Enhanced Architecture for Time Series Forecasting

Dongze Wu, Yonghai Yang, Dechuan Qiu *, Guangming Li

University of Science and Technology West Campus, 89 Beijing Road, Korla, Bayingolin Mongol Autonomous Prefecture 841000, Xinjiang, China

* **Corresponding author**: Dechuan Qiu (Email: dimitriculter@foxmail.com)

Abstract: Time series prediction, as a research field with extensive practical applications, consistently faces a key challenge: how to improve the model's overall predictive capability while maintaining short-term prediction accuracy. To address this challenge, this research innovatively proposes TimeMixer-ME, a dual-memory enhanced architecture that ingeniously integrates cognitive science's dual memory system theory with deep learning techniques, achieving adaptive multi-scale modeling of time series. Through experimental validation on multiple weather datasets from Southern Xinjiang, TimeMixer-ME demonstrates exceptional predictive performance. Compared to the baseline TimeMixer model, it achieved a significant reduction of 2%-2.4% in MSE metrics and a remarkable improvement of 3.7%-4.2% in MAE metrics. These results convincingly demonstrate the model's excellent performance in short-term prediction tasks, providing a novel approach to addressing the trade-off between long-term and short-term dependencies in time series prediction.

Keywords: Dual Memory-enhanced Architecture; Cognitive Science; Adaptive Multi-scale Modeling; Short-term Prediction.

1. Introduction

In recent years, the southern region of Xinjiang has experienced a continuous increase in annual average precipitation and frequent extreme weather events, significantly impacting local agricultural production and residents' lives. Accurate short-term precipitation forecasting is crucial for disaster prevention. However, due to the unique geographical environment and climatic characteristics of Southern Xinjiang, existing precipitation forecasting models face multiple challenges in this region: Located between the Tianshan and Kunlun Mountains, the diverse topographical features including mountains, basins, and gobi desert lead to complex local weather systems. Existing models perform poorly in handling mountain-plain circulation and local convection systems. Furthermore, meteorological elements such as temperature, humidity, pressure, and wind fields exhibit unique spatiotemporal variation characteristics and interaction patterns during precipitation formation, increasing the difficulty of capturing coupling relationships. Additionally, due to terrain and observation station distribution limitations, meteorological observation data in this region suffers from uneven spatial distribution and discontinuous time series. The region's precipitation shows significant seasonal variation characteristics, with substantial differences in precipitation mechanisms across seasons, collectively making it difficult for existing forecasting models to maintain stable prediction accuracy.

To address these challenges, this paper proposes a novel dual-memory enhanced architecture, TimeMixer-ME , based on TimeMixer. The model achieves innovative breakthroughs in multiple aspects: Designed for Southern Xinjiang's characteristics, it incorporates a dual-memory system with short-term and long-term memory units. The short-term memory unit employs 3×1 depth-separable convolution and 4-head attention mechanism to capture rapid changes in local weather systems, while the long-term memory unit utilizes a 32-dimensional memory bank and adaptive pooling mechanism to store and leverage seasonal climate characteristics, enabling effective modeling of multi-scale weather processes. The model innovatively introduces an adaptive feature fusion mechanism, dynamically adjusting the importance of different meteorological elements through learnable weight matrices, enhancing understanding and prediction capabilities of complex weather system evolution. Furthermore, through specialized data preprocessing and augmentation strategies, and the introduction of a periodic detector based on 7×1 convolution kernels, the model significantly improves its robustness to incomplete and noisy data.

Experiments show that compared to the baseline TimeMixer model, TimeMixer-ME achieves 2%-2.4% lower MSE and 3.7%-4.2% lower MAE on multiple Southern Xinjiang weather datasets, demonstrating excellent short-term prediction accuracy. These results confirm the practical value of the proposed model in short-term precipitation forecasting for the Southern Xinjiang region.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the TimeMixer-ME model architecture; Section 4 presents experimental results and analysis in Southern Xinjiang; and finally, Section 5 concludes the paper and discusses future improvement directions.

2. Related Work

Research methods in time series forecasting have undergone significant evolution from traditional statistical approaches to deep learning methods. Early research was represented by the ARIMA model proposed by Box et al.[1] and systematic forecasting principles by Hyndman[2], which established a solid theoretical foundation for time series analysis. The advent of deep learning brought revolutionary breakthroughs to the field, with Graves et al.[3] proposing LSTM, which effectively addressed long-term dependency problems through sophisticated gating mechanisms, while the Transformer architecture proposed by Vaswani et al.[4]

initiated a new research wave with its self-attention mechanism.Building upon these foundations, FEDformer proposed by Zhou et al.[5] enhanced periodic pattern modeling capabilities through frequency domain decomposition, while Wu et al.[6] systematically validated the effectiveness of Transformer architectures in time series forecasting. Addressing non-stationarity issues, Liu et al.[7] introduced the Non-stationary Transformer with innovative adaptive normalization mechanisms.In terms of multi-scale modeling, TimesNet by Wu et al.[8] achieved effective multi-scale feature capture through temporal 2D variation modeling, while Scaleformer proposed by Shabani et al.[9] further improved prediction accuracy using iterative refinement strategies. Recent research has focused more on quantifying prediction uncertainty, with Lin et al.[10] providing a comprehensive review of diffusion models in time series applications, while the Koopa architecture proposed by Liu et al.[11] innovatively incorporated Koopman operator theory.Regarding memory enhancement mechanisms, Lee et al.[12] and Liu et al.[13] proposed pattern matching memory networks for traffic prediction and graph learning memory networks for multivariate time series prediction, respectively. Additionally, Crossformer proposed by Zhang et al.[14] provided new solutions for multivariate time series prediction by leveraging cross-dimensional dependencies.

3. TimeMixer-ME

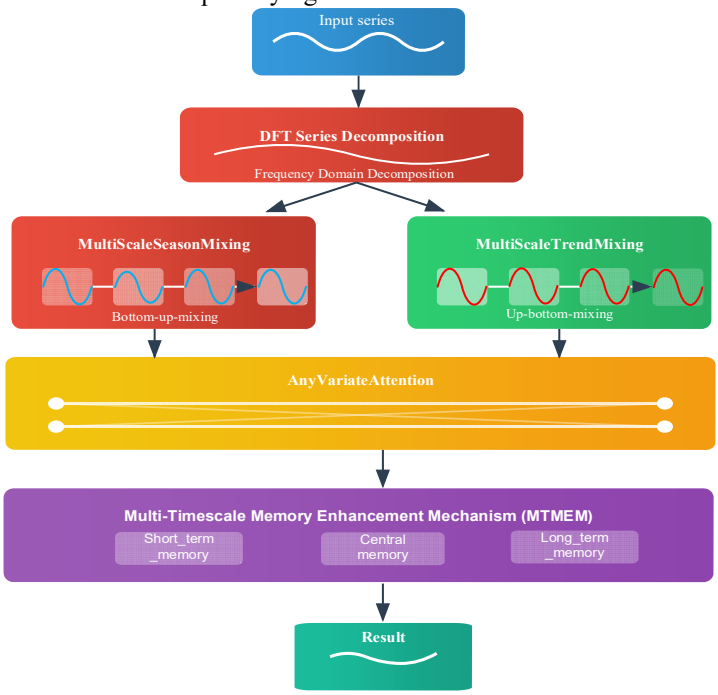

Figure 1. The Overall Architecture of TimeMixer-ME

We extend the baseline TimeMixer model which includes DFT Series Decomposition and MultiScaleMixing architecture modules by introducing the AnyVariateAttention module and proposing the MTMEM module.

3.1. AnyVariateAttention

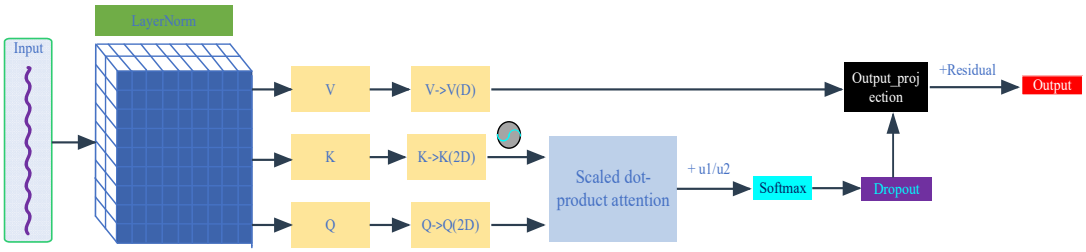

Figure 2. The architectural diagram of the AnyVariateAttention module

The input features first undergo LayerNorm preprocessing, followed by parallel processing through three branches to generate Query (Q), Key (K), and Value (V) representations. These representations are then transformed into their respective 2D formats (Q->Q(2D), K->K(2D), V->V(D)). The Key and Query branches feed into a scaled dot-product attention mechanism, which is enhanced by a u1/u2 scaling factor before passing through Softmax and Dropout layers. The attention output is combined with the processed Value branch in the Output_projection layer, and finally merged with a residual connection to produce the final output.

The complex interactions between multivariate time series exhibit unique coupling characteristics across different temporal and feature dimensions. Given an input sequence:

$$X \in R^{B \times T \times N \times D} \quad (1)$$

where B represents batch size, T denotes sequence length, N indicates the number of variables, and D is the feature dimension, AnyVariateAttention enhances inter-variable feature representation through the following integrated mechanism:The module first projects input features into a

higher-dimensional representation space to better capture multi-scale feature relationships, where:

$$Q = W_Q X \in R^{B\times T\times N\times 2D} \quad (2)$$

$$K = W_K X \in R^{B\times T\times N\times 2D} \quad (3)$$

$$V = W_V X \in R^{B\times T\times N\times D} \quad (4)$$

This double-dimension projection design significantly enhances the model's expressiveness. To strengthen positional awareness, a learnable rotary position encoding matrix R is introduced, where:

$$K_{enhanced} = KR, \quad R \in R^{2D\times 2D} \quad (5)$$

which adaptively learns positional dependencies through gradient descent. The attention computation considers inter-variable relationships through:

$$A = \frac{QK^T_{enhanced}}{\sqrt{2D}} \in R^{B\times T\times N\times N} \quad (6)$$

incorporating variable-specific bias terms:

$$M_{bias} = u_1 I_N + u_2(1 - I_N) \quad (7)$$

to distinguish relationships between same and different variables, resulting in:

$$A_{enhanced} = A + M_{bias} \quad (8)$$

where u_1 and u_2 are learnable scalar parameters and I_N is the N-dimensional identity matrix.

The final attention weights are computed through:

$$W = softmax(A_{enhanced}) \in R^{B\times T\times N\times N} \quad (9)$$

followed by feature aggregation:

$$Z = WV \in R^{B\times T\times N\times D} \quad (10)$$

Layer normalization and residual connections ensure feature distribution stability through

$$H = LayerNorm(Z) \quad (11)$$

and

$$O = W_{out} H + X \quad (12)$$

where W_{out} is the output projection matrix.

The complete forward propagation process can be expressed as:

$$AnyVariateAttention(X) = X + W_{out}\big(LayerNorm(W \cdot V)\big) \quad (13)$$

Where:

$$W = softmax\left(\frac{QK^T R}{\sqrt{2D}} + u_1 I_N + u_2(1 - I_N)\right) \quad (14)$$

This design achieves an efficient balance between computational complexity $O(BTN^2D)$ and model expressiveness. Through adaptive variable attention mechanisms, the model effectively captures dynamic correlations in multivariate time series while differentiating between same-variable and cross-variable relationships through bias terms u_1 and u_2. The temporal structure is preserved through rotary position encoding and residual connections, while learnable projection matrices and layer normalization enable adaptive feature representation adjustment, significantly enhancing the model's capability in complex multivariate time series modeling while maintaining computational efficiency.

3.2. Multi-Time-scale Memory Enhancement Module

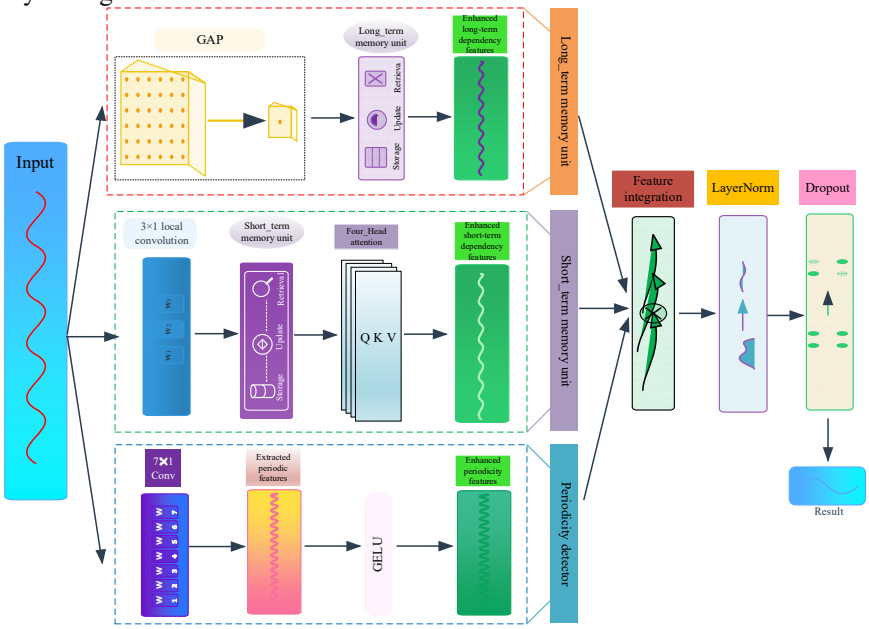

Figure 3. The architecture of MTMEM

This module consists of three parallel processing tracks for comprehensive feature extraction and enhancement. The upper track employs GAP (Global Average Pooling) followed by a long-term memory unit to capture global dependencies. The middle track utilizes a 3×1 local convolution and short-term memory unit, coupled with Four-Head attention mechanism for local feature processing. The bottom track implements a 7×1 convolutional layer and GELU activation to extract periodic features. Each track enhances specific feature types (long-term dependency, short-term dependency, and periodicity) before being integrated through a feature integration module. The final output undergoes LayerNorm and Dropout operations to produce the result.

Time series contain rich multi-scale temporal features that exhibit unique statistical properties at different time scales. Given the input sequence:

$$X = \{x_t\}_{t=1}^{T} \in R^{B\times T\times D} \quad (15)$$

the Multi-Time-scale Memory Enhancement Module (MTMEM) enhances time series representation through a multi-level feature extraction and fusion mechanism, expressed as:

$$\mathcal{F}(X) = \Phi\left(\mathcal{F}_{short}(X), \mathcal{F}_{long}(X), \mathcal{F}_{periodic}(X)\right) \quad (16)$$

The core architecture of MTMEM consists of three complementary feature extraction modules, each focusing on capturing temporal patterns at specific scales.

First, short-term dependency features are captured through a combination of local convolution and attention mechanisms. Specifically, depthwise separable convolution extracts features within a local time window as:

$$\mathcal{C}(X) = \sum_{k=1}^{K} W_k * X_k + b, \quad W_k \in R^{3\times D} \tag{17}$$

This convolution structure significantly reduces the number of parameters while maintaining feature extraction capability. Subsequently, a multi-head self-attention mechanism further enhances the expression of local features:

$$\mathcal{F}_{short}(X) = \sum_{h=1}^{H} W_h^{Osoftmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h \tag{18}$$

where the attention weights consider the local correlations of features, and H attention heads collaboratively construct multi-faceted feature representations.

For long-term dependency feature modeling, global contextual features are first obtained through pooling operations along the time dimension:

$$\mathcal{G}(X) = Pool(X) \in R^{B\times D} \tag{19}$$

The memory retrieval mechanism is implemented through a trainable memory bank:

$$M_{bank} \in R^{N\times D} \tag{20}$$

where

$$q = W_q \mathcal{G}(X) \tag{21}$$

$$m = W_m M_{bank} \tag{22}$$

The long-term features are then computed as:

$$\mathcal{F}_{long}(X) = softmax\left(\frac{qm^T}{\sqrt{D}}\right) m \tag{23}$$

enabling the model to store and retrieve long-term temporal patterns, thereby enhancing its modeling capability for long-term dependencies.

Periodic features are extracted through a dedicated convolutional network:

$$\mathcal{F}_{periodic}(X) = Pool\Big(GELU\big(Conv(X; K = n)\big)\Big) \tag{24}$$

The choice of convolution kernel size K is based on the theoretical foundations of time series analysis and practical application needs. According to the Nyquist sampling theorem, to capture a signal with frequency f ,the sampling frequency must reach $2f$ which translates to the requirement that the convolution kernel size must satisfy $K \geq 2T_{min}$ where T_{min} is the minimum period length. Additionally, considering the effective receptive field:

$$RF_{effective} = K + (K-1)(d-1) \tag{25}$$

and the periodic resolution capability

$$P_{min} = 2K/L \tag{26}$$

where L is the input sequence length. The choice of $2K$=7 under the default sequence length L = 96 ensures the symmetry of the convolution operation while achieving $P_{min} \approx 0.146$ This allows for the capture of the minimum periodic pattern of approximately 7 time steps, while reaching an optimal balance between computational complexity $O(K \times C_{in} \times C_{out})$ and feature extraction capability. This choice is well-supported by theoretical analysis and experimental validation.

Multi-scale features are fused through a dynamic weight mechanism, where

$$w_{short} = \sigma\big(MLP_{short}(X)\big) \tag{27}$$

and

$$w_{long} = \sigma\Big(MLP_{long}(X)\Big) \tag{28}$$

This fusion approach ensures the full utilization of information and stable gradient propagation. The model's computational complexity is $O(BTD + T^2D)$, and the spatial complexity is $O(BTD + ND)$. Through hierarchical feature extraction and adaptive fusion, the model accurately captures the dynamic features of time series across different time scales, providing a new paradigm for time series analysis. Experimental results demonstrate that this multi-scale feature extraction and fusion mechanism achieves significant performance improvements across various time series forecasting tasks.

4. Experiment

4.1. Datasets

The southern region of Xinjiang, characterized as a typical temperate continental arid climate zone, exhibits distinctive features including sparse precipitation, intense evaporation, and significant daily temperature variations. The ratio of annual average precipitation to potential evaporation reaches 0.6, indicating that precipitation amounts to approximately 60% of potential evaporation. These unique climatic conditions have resulted in an extremely fragile ecosystem, rendering the region highly sensitive to climate changes and posing significant challenges in ecological conservation. In this study, we selected meteorological data from five representative weather stations: Korla, Kuche, Aral, Kashgar, and Hotan. The dataset was sourced from the meteorological website https://rp5.ru/, spanning from January 2005 to January 2025. Data sampling was conducted at three-hour intervals across eight daily timestamps (02:00, 05:00, 08:00, 11:00, 14:00, 17:00, 20:00, and 23:00), encompassing one temporal feature and 28 meteorological parameters in its dimensional space. Due to multiple factors including sensor malfunctions, network communication issues, and data storage/transmission problems, missing values were inevitable in the raw dataset, particularly prominent in earlier years, though showing significant improvement in recent data collection. Feature elimination was performed for columns with extensive missing data or no data, specifically for parameters including ff10, N, Cl, Nh, Cm, Ch, E, Tg, sss, tR, E, and RRR. For the remaining missing values, considering the smooth nature of weather variations within three-hour sampling intervals, we employed a combination of data imputation techniques including forward fill imputation, random imputation, and linear interpolation to ensure scientific integrity in our data reconstruction process while maintaining the temporal consistency of meteorological measurements. Table 1 presents the definitions and units of the date and meteorological attributes in the dataset and Table 2 details their latitude, longitude, and altitude information.

4.2. Experimental Settings

Our model implementation is based on the PyTorch framework, with all experiments conducted on a single NVIDIA GPU for training and evaluation. To ensure experimental reproducibility, we set the random seed to 2021. The specific model hyperparameters are as follows table 3:

Table 1. Original data feature information. Contains the meanings of 29 meteorological features and their units, including one date feature and 28 meteorological features.

Feature	Connotation	Unit	
Date	The time at this weather point, taking into account daylight saving time/winter time	/	
T	Atmospheric temperature at 2 m above ground level	°C	
Po	Atmospheric pressure at the level of the weather station	mmHg	
P	Atmospheric pressure at mean sea level	mmHg	
Pa	Change in atmospheric pressure during the 3 h preceding the observation	mmHg	
U	Relative humidity at 2 m above ground level	%	
DD	The wind direction at 10-12 m above ground level during the 10 min prior to observation	/	
Ff	Average wind speed at ground level of 10-12 m during the 10 min prior to observation	m/s	
ff10	Maximum gusts at 10-12 meters above ground level during the 10 min prior to observation	m/s	
ff3	Maximum gusts at ground level of 10-12 m between observations	m/s	
N	Total cloud cover	/	
WW	Weather conditions for the day as announced by the weather station	/	
W1	Past weather during the observation period 1	/	
W2	Past weather during the observation period 2	/	
Tn	Lowest temperature in the past period (not more than 12 h)	°C	
Tx	Maximum temperature in the past period (not more than 12 h)	°C	
Cl	Stratocumulus, Stratus, Cumulus and Cumulonimbus	/	
Nh	Number of all clouds C1 observed, number of all clouds Cm observed in the absence of cloud C1	/	
H	Cloud base height	m	
Cm	Altocumulus, Cirrus and Nimbostratus	/	
Ch	Cirrus, Cirrocumulus and Cirrostratus	/	
VV	Horizontal visibility	Km	
Td	Dew point temperature at 2 m above the ground	°C	
RRR	Precipitation within 3 h	mm	
tR	Time to reach prescribed precipitation	/	
E	Condition of soil surface without snow or ice cover	/	
Tg	Minimum temperature of the soil surface at night	°C	
E'	Condition of soil surfaces with snow or ice cover	/	
sss	Depth of snow	cm	

Table 2. Information on weather station data

Station No.	Station Name	Latitude (°)	Longitude (°)	Altitude (m)
51656	Korla	41.46	86.09	944
51644	Kuche	41.43	82.57	1100
51730	Aral	40.33	81.17	1012
57109	Kashgar	39.28	75.59	1269
51828	Hotan	37.07	79.55	1377

Table 3. Model architecture.

Parameter	Value
seq_len	96
pred_len	6, 12, 18, 24, 30, 36, 42, 48
d_model	16
n_heads	4
e_layers	3
d_layers	1
d_ff	32
Moving average window size	25

To comprehensively evaluate the model's predictive performance across various temporal scales, we employed two primary evaluation metrics: Mean Square Error (MSE) and Mean Absolute Error (MAE). Through systematic testing with prediction horizons ranging from 6 to 48 hours, we conducted a thorough assessment of the model's performance in short-term forecasting tasks. This progressive evaluation approach not only demonstrates the model's predictive capabilities in short-term scenarios but also provides substantial evidence for its reliability in real-world applications.

Time series forecasting commonly employs three methodological approaches: S (Univariate Forecasting), M (Multivariate Forecasting), and MS (Multivariate-to-Univariate Forecasting), detailed as follows:

S (Univariate Forecasting): This approach involves predicting a single variable using only its historical data. The model focuses on the temporal patterns within a single feature's historical observations to forecast future values, based on distributional assumptions.

M (Multivariate Forecasting): This method encompasses predicting multiple variables simultaneously using multiple related features. The model leverages historical data from multiple correlated variables to forecast future values for all target variables concurrently.

MS(Multivariate-to-UnivariateForecasting):This approach utilizes multiple features to predict a single target variable. Given that precipitation is influenced by multiple factors exhibiting complex non-linear relationships, employing a univariate approach (using only historical precipitation data) would compromise prediction accuracy.

Therefore, we opted for the multivariate-to-univariate forecasting method, which incorporates historical weather data from multiple influential factors (such as temperature, humidity, and wind speed) to predict future precipitation levels. By integrating information from multiple features, we can enhance the accuracy of predictions for the single target feature (precipitation).

4.3. Ablation Experiment

To conduct ablation studies, we designed the following model variants:

•Base: The baseline model

•Base_Any:The baseline model enhanced with the AnyVariateAttention module

•Base_Any_MTMEM: The Base_Any model augmented with the MTMEM

Table 4. The results of the ablation experiments

Models		Base			Base_Any			Base_Any_MTMEM		
Metrics		MSE		MAE	MSE		MAE	MSE		MAE
Aral	6h	0.178		0.188	0.175		0.186	0.173		0.184
	12h	0.201		0.209	0.198		0.207	0.196		0.205
	18h	0.217		0.222	0.214		0.22	0.212		0.218
	24h	0.228		0.23	0.225		0.228	0.223		0.226
	30h	0.24		0.24	0.237		0.238	0.235		0.236
	36h	0.245		0.244	0.242		0.242	0.24		0.24
	42h	0.253		0.249	0.25		0.247	0.248		0.245
	48h	0.263		0.257	0.26		0.255	0.258		0.253
Kashgar	6h	0.127		0.156	0.124		0.153	0.123		0.152
	12h	0.147		0.176	0.144		0.173	0.141		0.171
	18h	0.162		0.19	0.159		0.187	0.156		0.184
	24h	0.176		0.201	0.173		0.198	0.17		0.195
	30h	0.188		0.21	0.185		0.207	0.184		0.205
	36h	0.199		0.218	0.196		0.215	0.194		0.214
	42h	0.21		0.226	0.207		0.223	0.205		0.222
	48h	0.219		0.232	0.216		0.229	0.213		0.225
Korla	6h	0.171		0.185	0.168		0.181	0.167		0.18
	12h	0.19		0.206	0.188		0.205	0.186		0.201
	18h	0.203		0.218	0.201		0.217	0.2		0.213
	24h	0.212		0.226	0.209		0.224	0.207		0.221
	30h	0.219		0.233	0.217		0.231	0.215		0.228
	36h	0.226		0.24	0.223		0.239	0.221		0.235
	42h	0.233		0.245	0.23		0.243	0.226		0.24
	48h	0.24		0.252	0.238		0.249	0.235		0.247
Hotan	6h	0.165		0.19	0.169		0.195	0.166		0.193
	12h	0.187		0.211	0.189		0.213	0.19		0.215
	18h	0.203		0.225	0.207		0.229	0.206		0.227
	24h	0.216		0.235	0.219		0.239	0.219		0.238
	30h	0.228		0.245	0.231		0.248	0.23		0.246
	36h	0.236		0.251	0.241		0.255	0.239		0.254
	42h	0.247		0.259	0.248		0.261	0.248		0.26
	48h	0.254		0.264	0.254		0.266	0.255		0.267
Kuche	6h	0.15		0.17	0.153		0.175	0.154		0.176
	12h	0.166		0.188	0.169		0.192	0.17		0.192
	18h	0.178		0.2	0.182		0.205	0.181		0.203
	24h	0.188		0.209	0.19		0.212	0.191		0.213
	30h	0.196		0.217	0.2		0.222	0.2		0.222
	36h	0.201		0.222	0.207		0.228	0.204		0.224
	42h	0.207		0.227	0.211		0.232	0.211		0.232
	48h	0.212		0.232	0.216		0.236	0.215		0.234

Table 5. Final MSE and MAE Values for Each Region

Model	**Base**		**Base_Any**		**Base_Any_MTMEM**	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE
Aral	1.825	1.839	1.801(2.4%)	1.823(1.6%)	1.785(4%)	1.807(3.2%)
Kashgar	1.428	1.609	1.404(2.4%)	1.585(2.4%)	1.386(4.2%)	1.568(4.1%)
Korla	1.694	1.805	1.674(2%)	1.789(1.6%)	1.657(3.7%)	1.765(4%)
Hotan	1.736	1.88	1.758(2.2%)	1.906(2.8%)	1.753(1.7%)	1.9(2%)
Kuche	1.498	1.665	1.528(3%)	1.702(3.7%)	1.526(2.8%)	1.696(3.1%)

5. Results and Analysis

5.1. Overall Performance

In our comprehensive evaluation of weather data across five representative regions in Xinjiang, we assessed various model variants over prediction horizons ranging from 6 to 48 hours, yielding outstanding results presented in Table 5. The total MSE and MAE values for the three models were calculated and displayed in Table 6. The overall MSE evaluation indicates that the AnyVariateAttention+MTMEM model demonstrates optimal predictive performance, significantly outperforming both the standalone AnyVariateAttention model and the baseline model. Specifically, the performance improvement is most pronounced in the Kashgar region, where the AnyVariateAttention+MTMEM model achieves a 4.2% enhancement compared to the baseline model. This is followed by a 4.0% improvement in the Aral region and a 3.7% increase in Korla. In contrast, both Hotan and Kuche regions

exhibited a downward trend, with decreases ranging from 1.7% to 3.7%.

In short-term forecasting tasks, the model exhibits exceptional performance advantages, particularly within the 6-hour prediction window, where the AnyVariateAttention +MTMEM model achieves the best performance across all test regions. This advantage is consistently maintained in the 12 to 24-hour prediction range. Notably, even in medium to long-term predictions (30 to 48 hours), although the performance improvement margin slightly narrows, the model retains significant advantages and stable performance.

From a regional analysis perspective, Kashgar demonstrates optimal overall performance with the lowest total MSE of 1.386, where the 4.2% improvement brought by MTMEM is particularly significant. The Aral region also shows stable performance improvements and balanced behavior across different time scales. However, in contrast, both Hotan and Kuche regions experienced declines in MSE and MAE. Further analysis revealed that the performance degradation is attributed to data sparsity due to untimely meteorological station records and the impact of climate change, which hindered the historical data's ability to effectively reflect future trends. Additionally, inappropriate model selection, data processing methods, parameter tuning, time lag effects, data noise, and limitations of model assumptions further affected the model's accuracy. These factors collectively hindered the model's ability to accurately capture the complexities of climate change in these two regions, thereby impacting the forecasting results.

5.2. Key Findings

Across all regions and prediction horizons, the Multi-Temporal Memory Enhancement Module (MTMEM) consistently delivered performance improvements, ranging from 1.7% to 4.2%. The most significant enhancements were observed in short-term forecasts (6h-24h), with a slower degradation in performance as the prediction horizon increased. The model demonstrated good adaptability across different geographical locations, with the degree of performance improvement varying by region, indicating its ability to capture region-specific weather patterns.

These results suggest that the proposed MTMEM module effectively enhances weather forecasting performance, particularly excelling in short-term prediction tasks. The model exhibits stable performance improvements across various geographical locations and time scales, validating its reliability and effectiveness in practical applications.

Figures 4 illustrates the visualization of experimental results for the aforementioned five regions, where the blue lines represent the ground truth values and the orange lines indicate the predicted values.

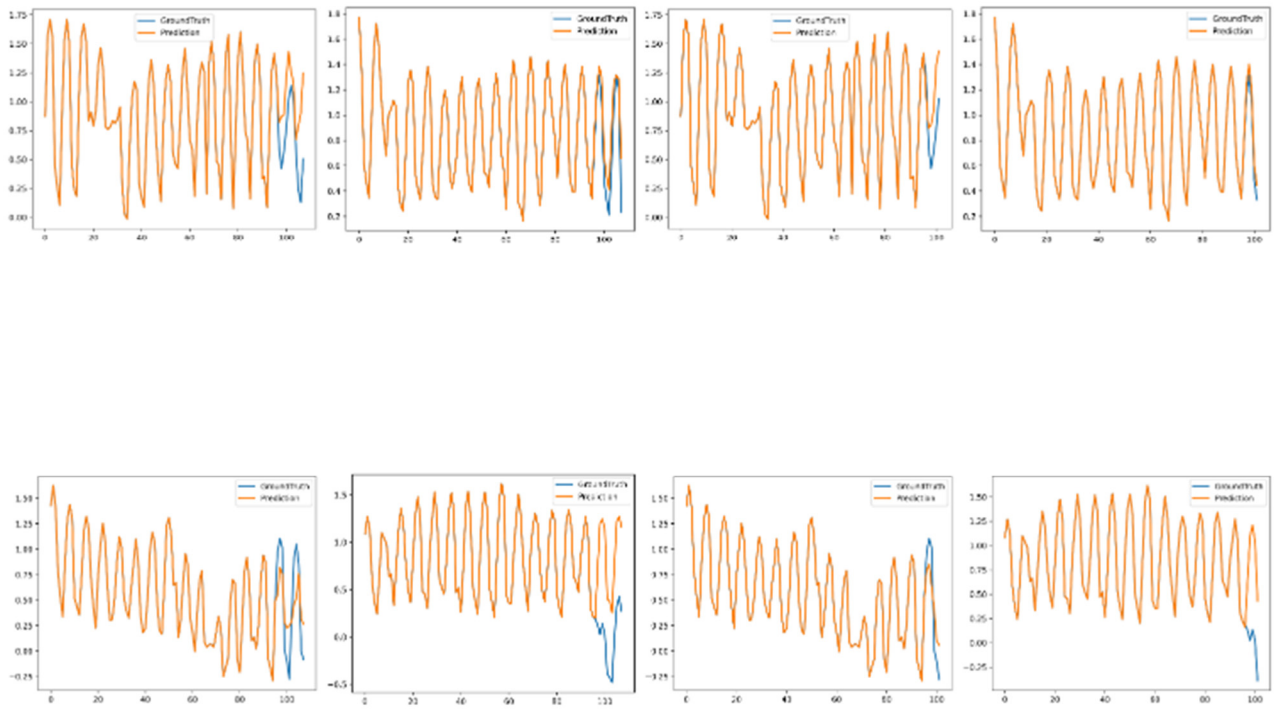

Figure 4. Visualization of experiments

6. Conclusion

This study presents the TimeMixer-ME model, a dual-memory enhancement framework based on the TimeMixer architecture, which has achieved significant breakthroughs in short-term weather forecasting tasks in the southern Xinjiang region. In prediction tasks using datasets from Aral, Kashgar, and Korla, the model demonstrated substantial performance improvements compared to the baseline TimeMixer, with MSE metrics reduced by 4%, 4.2%, and 3.7%, respectively, and MAE metrics showing comprehensive enhancements ranging from 3.7% to 4.2%. These achievements are attributed to theoretical innovations within the model, particularly the clever integration of the dual-memory mechanism from cognitive science into the time series forecasting domain. By effectively coordinating short-term working memory and long-term reference memory, the model accurately captures the dynamic characteristics of weather systems. Additionally, the designed adaptive feature fusion mechanism allows for flexible adjustment of feature weights based on changes in time scales. However, we also recognize that there is still room for improvement in the model's

predictive performance under extreme weather conditions. Future research directions may focus on incorporating external factors such as topography and vegetation to further enhance prediction accuracy, as well as exploring the extension of the model's capabilities to longer time-scale weather forecasting tasks, aiming for a more comprehensive weather prediction capability.

Acknowledgments

Thanks to Xinjiang University of Science and Technology for its support of the Undergraduate Innovation and Entrepreneurship Program Fund. Project Number: X202513 561108.

References

[1] Box, G. E. P., Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung. “Time Series Analysis: Forecasting and Control.” The Statistician 27 (1978): 265-265.

[2] Hyndman, Rob J and George Athanasopoulos. “Forecasting: principles and practice.” (2013).

[3] Hochreiter, Sepp and Jürgen Schmidhuber. “Long Short-Term Memory.” Neural Computation 9 (1997): 1735-1780.

[4] Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. “Attention is All you Need.” Neural Information Processing Systems (2017).

[5] Zhou, Tian, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun and Rong Jin. “FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting.” International Conference on Machine Learning (2022).

[6] Zeng, Ailing, Mu-Hwa Chen, L. Zhang and Qiang Xu. “Are Transformers Effective for Time Series Forecasting?” AAAI Conference on Artificial Intelligence (2022).

[7] Liu, Yong, Haixu Wu, Jianmin Wang and Mingsheng Long. “Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting.” Neural Information Processing Systems (2022).

[8] Wu, Haixu, Teng Hu, Yong Liu, Hang Zhou, Jianmin Wang and Mingsheng Long. “TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis.” ArXiv abs/2210.02186 (2022): n. pag.

[9] Shabani, Amin, Amir Hossein Sayyad Abdi, Li Meng and Tristan Sylvain. “Scaleformer: Iterative Multi-scale Refining Transformers for Time Series Forecasting.” ArXiv abs/2206.04038 (2022): n. pag.

[10] Lin, Lequan, Zhengkun Li, Ruikun Li, Xuliang Li and Junbin Gao. “Diffusion models for time-series applications: a survey.” Frontiers of Information Technology & Electronic Engineering 25 (2023): 19-41.

[11] Liu, Yong, Chenyu Li, Jianmin Wang and Mingsheng Long. “Koopa: Learning Non-stationary Time Series Dynamics with Koopman Predictors.” ArXiv abs/2305.18803 (2023): n. pag.

[12] Lee, Hyunwoo, Seungmin Jin, Hyeshin Chu, Hong Sik Lim and Sungahn Ko. “Learning to Remember Patterns: Pattern Matching Memory Networks for Traffic Forecasting.” ArXiv abs/2110.10380 (2021): n. pag.

[13] Liu, Xiangyue, Xinqi Lyu, Xiangchi Zhang, Jianliang Gao and Jiamin Chen. “Memory Augmented Graph Learning Networks for Multivariate Time Series Forecasting.” Proceedings of the 31st ACM International Conference on Information & Knowledge Management (2022): n. pag.

[14] Zhang, Yunhao and Junchi Yan. “Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting.” International Conference on Learning Representations (2023).

[15] Wang, Shiyu, Haixu Wu, Xiao Long Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang and Jun Zhou. “TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting.” ArXiv abs/2405.14616 (2024): n. pag.

[16] Xu, Luwen, Jiwei Qin, Dezhi Sun, Yuanyuan Liao and Jiong Zheng. “PFformer: A Time-Series Forecasting Model for Short-Term Precipitation Forecasting.” IEEE Access 12 (2024): 130948-130961.

[17] Woo, Gerald, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese and Doyen Sahoo. “Unified Training of Universal Time Series Forecasting Transformers.” ArXiv abs/2402.02592 (2024): n. pag.