STGFM: A SPATIO-TEMPORAL GLOBAL FACTOR MODEL FOR MULTI-SERIES ELECTRICITY LOAD FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate multi-series multi-step load forecasting requires exploiting shared dynamics across series while preserving series-specific heterogeneity. However, cross-series associations evolve across forecasting instances, making fixed shared representations insufficient for these dual objectives. We propose a spatio-temporal global factor model for multi-series load forecasting, termed STGFM. A cross-series Transformer derives input-conditioned entmax aggregation weights that project historical loads onto compact time-indexed factors. A Mamba state-space module encodes historical factor dynamics, whose final representation is combined with known future covariates to generate multi-step factor forecasts. A conditional decoder then reconstructs series-specific trajectories from the predicted factors, series context, and forecast-step information. Experiments on the Real-world Load Dataset and the UCL Electricity Dataset show that STGFM reduces forecasting errors across multiple horizons. On the Real-world Load Dataset, STGFM reduces RMSE by 2.0%–7.0% across four horizons (12, 24, 48, and 96 hours) compared with TimesNet. On the UCL Electricity Dataset, the corresponding reductions are 8.2%–21.3% at 24, 48, and 96 hours relative to the strongest baseline at each horizon. Ablation studies further show that conditional reconstruction consistently reduces both RMSE and MAE compared with shared MLP decoding across both datasets and all four horizons.

1 INTRODUCTION

Driven by the rapid evolution of modern power systems, electricity consumption patterns have become increasingly diverse across series, making multi-series multi-step load forecasting a critical foundation for secure system operation and economic dispatch Liu et al. (2026). Load series exhibit shared temporal patterns arising from common driving factors such as calendar effects and meteorological conditions, while simultaneously retaining heterogeneity due to variations in consumption behavior, load composition, and demand response characteristics. Modeling each series independently fails to exploit cross-series information, whereas unified modeling with shared parameters risks overlooking series-specific characteristics. This fundamental tension between capturing shared dynamics and preserving individual heterogeneity poses a central challenge for multi-series load forecasting.

Existing approaches to multi-series time series forecasting largely follow two paradigms, namely channel independence and channel mixing. The former processes historical observations of each series independently and extracts temporal features through architectures such as recurrent networks, temporal convolution, or patch-wise self-attention Kong et al. (2025); Luo & Wang (2024). In the context of load forecasting, this line of work is further enriched by incorporating meteorological lag effects, production conditions, and other domain-specific features to improve predictive representations Ahmad et al. (2025); Morais et al. (2023); Gomez et al. (2025). While such methods preserve series-specific characteristics, the lack of explicit cross-series interaction prevents them from fully exploiting complementary information among correlated series Yu et al. (2024); Zhou & Wang (2025). Channel-mixing methods, by contrast, leverage shared information through cross-channel attention or global latent factor modeling Liu et al. (2024); Wang et al. (2024). Dynamic attention mechanisms primarily regulate information exchange among series, but the extraction of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

co-movements across multiple series and their subsequent evolution are often not explicitly modeled. Factor models address this by representing co-movements as low-dimensional time series, thereby enabling explicit prediction of shared dynamics. However, fixed factor loadings cannot readily accommodate time-varying cross-series associations, and without an output mapping conditioned on both series identity and prediction horizon, these models struggle to capture how identical shared dynamics manifest differently across series and time steps.

To address these limitations, we propose the Spatio-Temporal Global Factor Model (STGFM) for multi-series multi-step load forecasting. STGFM jointly models cross-series shared dynamics and series heterogeneity by adaptively constructing global factors that preserve temporal structure from cross-series contexts, encoding historical factor dynamics through Mamba Wang et al. (2025), and integrating known future temporal covariates to generate multi-step factor forecasts. A conditional reconstruction decoder then transforms the global factors into series-specific load forecasts by incorporating series context and prediction-horizon information. Because the historical context participates in both factor construction and conditional reconstruction, the extraction of shared dynamics and the series-specific mapping are learned jointly within a unified end-to-end framework. Experiments on the Real-world Load Dataset and the UCL Electricity Dataset show consistent gains in RMSE and R^2 relative to the evaluated learned baselines, particularly at medium and long horizons. Ablation studies reveal that conditional reconstruction consistently yields lower RMSE and MAE than a shared MLP decoder. The main contributions of this paper are summarized as follows:

1. We propose the Spatio-Temporal Global Factor Model (STGFM) for multi-series multi-step load forecasting. STGFM represents multi-series co-movements as low-dimensional global factors while preserving temporal structure, forecasts their evolution using historical dynamics and known future temporal covariates, and generates series-specific forecasts through conditional reconstruction, thereby jointly modeling shared dynamics prediction and series-specific mapping.
2. We develop a factor aggregation and conditional reconstruction mechanism jointly driven by historical context. The aggregation weights adapt to cross-series associations within each window, enabling the global factors to capture their variation across windows. Conditional reconstruction then modulates future factors using series characteristics and prediction-horizon information, thereby preserving series-specific load patterns at each future time step.
3. Extensive comparative experiments against 13 baseline methods across four prediction horizons on the Real-world Load Dataset and the UCL Electricity Dataset show consistent gains in RMSE and coefficient of determination (R^2) relative to the evaluated learned baselines. Ablation studies further show that component contributions depend on the dataset and forecasting horizon, while series-adaptive reconstruction exhibits the most consistent effect.

2 Related Work

Load forecasting requires the modeling of temporal dependencies, exogenous influences, and cross-series relationships among electricity consumers. Existing research has addressed such issues along three complementary directions: deep sequence modeling with covariate integration, cross-variate dependency modeling, and global factor-based forecasting. Relevant approaches within the three directions are reviewed below, with particular emphasis on global factor methods.

Deep Learning for Load Forecasting: Deep learning approaches to load forecasting focus on capturing temporal dynamics and modeling their interactions with exogenous variables. In multivariate time series forecasting, TimesNet Wu et al. (2023) models multi-periodic variations through two-dimensional temporal representations, whereas PatchTST Nie et al. (2023) employs temporal patches to capture local structures for long-term forecasting. The two frameworks offer representative paradigms for modeling periodic patterns and local temporal variations relevant to load forecasting.

While such frameworks emphasize temporal representation, load-specific studies further incorporate meteorological, production, and operational variables. Pentsos et al. Pentsos et al. (2025) combine LSTM-based recurrent modeling with Transformer attention for multi-step load forecasting. Niu et

al. Niu et al. (2026) model dynamic lag relationships between meteorological variables and load responses through lag alignment and temporal–channel interaction. For industrial load forecasting, Wu et al. Wu et al. (2025) integrate meteorological, production, and efficiency-related variables using temporal structure-preserving modeling and gated feature fusion. Collectively, the reviewed studies focus primarily on temporal modeling and the incorporation of exogenous variables. However, beyond temporal and exogenous-variable modeling, cross-series information in multi-series settings offers an additional source of predictive context that is not captured by the historical observations and covariates of each target series.

Cross-Variate Dependency Modeling: In multi-series forecasting settings, cross-variate modeling methods leverage cross-series information from auxiliary variables while preserving the inherent temporal characteristics of each individual sequence. Crossformer Zhang & Yan (2023) models inter-variable dependencies through a cross-dimension attention mechanism. DUET Qiu et al. (2025) combines temporal clustering with soft channel clustering to identify associations under heterogeneous patterns, and applies sparsification to restrict channel-wise interactions. Adapformer Luo et al. (2026) performs adaptive channel management during both encoding and decoding stages to select relevant variables for representation learning and prediction. Overall, these three works adopt distinct architectural paradigms for cross-variate interaction: Crossformer relies on full cross-dimension attention, whereas DUET and Adapformer explicitly regulate variable contributions to forecasting via sparsification constraints and adaptive channel selection, respectively.

A complementary line of research organizes cross-variate information through parallel representation schemes or relational structural modeling. FusionTimePatch Zhang et al. (2026) integrates channel-independent and channel-mixing representations to jointly capture local variations, global trends, and inter-variable dependencies. MTGNN Wu et al. (2020) learns directed relational dependencies among time series via adaptive graph learning. Despite their different formulation foundations, both approaches realize effective cross-series information aggregation: FusionTimePatch organizes cross-variate coupling through hierarchical feature interaction, while MTGNN achieves information transfer through relational propagation on learned graphs.

Global and Factor-Based Forecasting Global forecasting methods learn transferable patterns by sharing model parameters across multiple related series. DeepAR Salinas et al. (2020) trains a shared autoregressive network to generate probabilistic forecasts for individual series. Temporal Fusion Transformer Lim et al. (2021) integrates static attributes, historical inputs, and known future covariates for multi-horizon forecasting. Parameter sharing enables both models to exploit statistical regularities across series, while common temporal variation remains implicitly encoded in the forecasting function rather than explicitly represented as a time-indexed factor sequence.

Factor-based models explicitly represent shared temporal dynamics while retaining series-specific components. Deep Factors Wang et al. (2019) combines global deep factors with local random effects to capture common temporal structure and series-level uncertainty. DeepGLO Sen et al. (2019) extracts global structure through matrix factorization regularized by a temporal convolutional network and incorporates a local forecasting network to model individual series and associated covariates. SOFTS Han et al. (2024) aggregates channel representations into a centralized core and redistributes the aggregated representation to individual channels for feature fusion. Collectively, prior approaches model shared information through factor decomposition, coordinated global and local forecasting, or centralized representation aggregation, while preserving mechanisms for series-specific prediction.

In multi-series load forecasting, cross-series associations may change across historical windows, and shared dynamics may produce different responses across series and forecast steps. STGFM models both characteristics through input-conditioned factor construction and series-conditioned reconstruction. Cross-series context within each historical window determines factor aggregation weights, which remain shared across time steps in the same window and transform historical loads into a time-indexed factor sequence. Mamba encodes the temporal evolution of the factor sequence, and known future temporal covariates are incorporated to generate multi-step factor forecasts. A conditional reconstruction decoder maps predicted factors to series-specific load trajectories by integrating contextualized series representations, series-level statistical descriptors, future temporal covariates, and forecast-step information. Overall, STGFM formulates shared dynamics as adaptive time-indexed factors, predicts their temporal evolution, and reconstructs series-specific trajectories within a unified forecasting framework.

3 METHOD

3.1 OVERALL ARCHITECTURE

STGFM addresses multi-series multi-step forecasting over N load series. For each forecasting instance, the model takes as input a historical load window $\mathbf{X}^h \in \mathbb{R}^{B\times L\times N}$, the corresponding observation mask $\mathbf{M}^h \in \{0,1\}^{B\times L\times N}$, historical temporal covariates $\mathbf{C}^h \in \mathbb{R}^{B\times L\times C}$, future temporal covariates $\mathbf{C}^f \in \mathbb{R}^{B\times H\times C}$, and series-specific static features $\mathbf{S} \in \mathbb{R}^{N\times D_s}$. Here, B denotes the batch size, L the historical window length, H the forecasting horizon, C the temporal covariate dimension, and D_s the static feature dimension. The ground-truth future loads and corresponding forecasts are denoted by $\mathbf{Y}, \widehat{\mathbf{Y}} \in \mathbb{R}^{B\times H\times N}$.

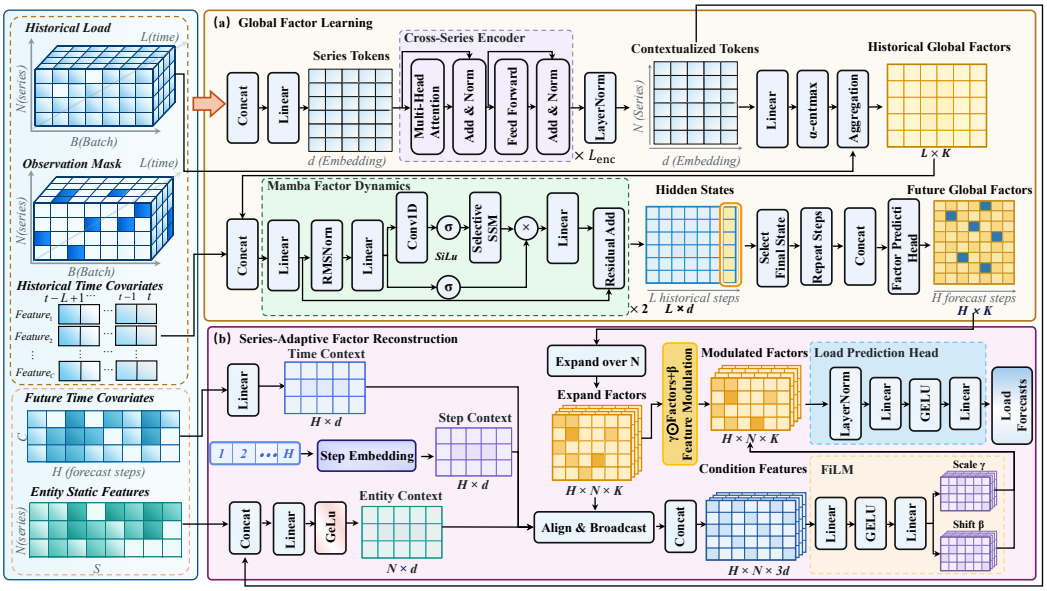


Figure 1: The overall architecture of STGFM. Historical multi-series inputs are compressed into global factors whose future trajectories are forecast using known temporal covariates, and the predicted factors are subsequently adapted to individual series to produce multi-step load forecasts.

As illustrated in Figure 1, STGFM follows a two-stage architecture comprising global factor construction and dynamics forecasting, followed by series-adaptive reconstruction. In the first stage, a cross-series Transformer encodes the historical observations, observation mask, and historical temporal covariates. Input-conditioned entmax aggregation then compresses the cross-series dimension into K global factors, represented as $\mathbf{F}^h \in \mathbb{R}^{B\times L\times K}$ with $K < N$. A Mamba-based state-space module models the historical factor dynamics and predicts the future factor trajectories $\widehat{\mathbf{F}}^f \in \mathbb{R}^{B\times H\times K}$, conditioned on the known future temporal covariates. In the second stage, the predicted factors are expanded across series and combined with future temporal information, forecast-step information, and series-specific static features. A feature modulation layer adapts the shared factor representation to each series and forecasting step, after which a load prediction head produces the final multi-series forecasts $\widehat{\mathbf{Y}}$.

3.2 GLOBAL FACTOR LEARNING AND DYNAMICS

Rather than forecasting directly in the N-dimensional series space, STGFM first represents cross-series shared variation using a compact set of time-indexed global factors. The factor composition is conditioned on the current forecasting instance, whereas the temporal evolution is modeled after the series dimension has been compressed. This separates input-dependent cross-series aggregation from factor-space dynamics forecasting while retaining the historical time axis.

For each series, the historical normalized load vector and its observation-mask vector are concatenated. The historical temporal covariates are summarized by their mean over the L historical steps and by the covariate vector at the final historical step:

$$\mathbf{c}_b^h = \mathrm{Concat}\left(\frac{1}{L}\sum_{t=1}^{L}\mathbf{C}_{b,t,:}^h, \mathbf{C}_{b,L,:}^h\right). \tag{1}$$

The summary $\mathbf{c}_b^h$ is shared across the N series in the same forecasting instance. For series n, the historical input representation is constructed as

$$\mathbf{r}_{b,n} = \mathrm{Concat}\left(\mathbf{X}_{b,:,n}^h, \mathbf{M}_{b,:,n}^h, \mathbf{c}_b^h\right). \tag{2}$$

Here, missing historical load values have been replaced by the corresponding series training mean before entering the model, while $\mathbf{M}^h$ retains the original observation validity. A token projection maps the series-wise representations to the model dimension, and a Cross-Series Transformer processes the resulting tokens along the series dimension:

$$\mathbf{Z} = \mathrm{CST}\left(\phi_{\mathrm{tok}}\left(\{\mathbf{r}_{b,n}\}_{n=1}^N\right)\right), \tag{3}$$

where $\mathbf{Z} \in \mathbb{R}^{B\times N\times D_z}$ denotes the contextualized series representations. The attention axis is the series dimension, so the Transformer provides cross-series context for the subsequent factor construction rather than modeling the historical time positions directly.

The contextualized representations parameterize K factor-specific aggregation patterns. A linear projection produces aggregation scores $\mathbf{U} \in \mathbb{R}^{B\times N\times K}$:

$$\mathbf{U} = \phi_{\mathrm{fac}}(\mathbf{Z}). \tag{4}$$

For each factor, α-entmax is applied over the N series. The parameter α is treated as a fixed hyperparameter of the selected model configuration:

$$A_{b,n,k} = \mathrm{entmax}_\alpha\left(\mathbf{U}_{b,:,k}\right)_n. \tag{5}$$

The resulting weights $\mathbf{A} \in \mathbb{R}^{B\times N\times K}$ are nonnegative, normalized over the series dimension, and may contain exact zeros. Within one forecasting instance, the same weights are shared across the L historical positions. The k-th global factor at historical position t is therefore obtained by

$$F_{b,t,k}^h = \sum_{n=1}^{N} A_{b,n,k} X_{b,t,n}^h. \tag{6}$$

This operation reduces the series dimension from N to K while preserving the L historical time positions, yielding $\mathbf{F}^h \in \mathbb{R}^{B\times L\times K}$ with $K < N$. Thus, the output consists of time-indexed factor trajectories rather than a static latent representation.

The factor trajectories are then propagated along the historical time axis. A projection ϕ_h maps the concatenation of historical factors and historical temporal covariates to the input dimension of the Mamba module:

$$\mathbf{R}^h = \phi_h\left(\mathrm{Concat}\left(\mathbf{F}^h, \mathbf{C}^h\right)\right). \tag{7}$$

The Mamba module produces a representation at each historical position:

$$\mathbf{H}^h = \mathrm{Mamba}\left(\mathbf{R}^h\right). \tag{8}$$

The final historical representation is repeated over the H future positions. In parallel, a separate projection ϕ_f maps the future temporal covariates to the factor forecasting space:

$$\mathbf{Q}^f = \mathrm{Repeat}_H\left(\mathbf{H}_{:,L,:}^h\right), \qquad \mathbf{E}^f = \phi_f(\mathbf{C}^f). \tag{9}$$

The factor prediction head combines the repeated historical representation with the step-specific future temporal representation:

$$\widehat{\mathbf{F}}^f = g_{\mathrm{factor}}\left(\mathrm{Concat}\left(\mathbf{Q}^f, \mathbf{E}^f\right)\right). \tag{10}$$

The resulting tensor $\widehat{\mathbf{F}}^f \in \mathbb{R}^{B\times H\times K}$ contains the predicted trajectories of the global factors. The historical context is shared across future positions, whereas the projected future temporal covariates distinguish the individual forecasting steps.

3.3 SERIES-ADAPTIVE FACTOR RECONSTRUCTION

The predicted global factors describe dynamics shared across series but do not by themselves specify the response of an individual series at a particular forecasting step. Series-Adaptive Factor Reconstruction introduces series- and horizon-specific information through conditional feature modulation. The reconstruction condition combines future temporal information, forecast-step information, and a series-specific context derived from both historical cross-series encoding and static series features.

The future factor tensor $\widehat{\mathbf{F}}^f$ is expanded along the series dimension and aligned with the N load series. Future temporal covariates are projected into temporal context vectors, and the forecast indices are mapped to forecast-step embeddings. The contextualized representation $\mathbf{Z}$ and the static features $\mathbf{S}$ jointly define the series-specific context:

$$\mathbf{e}^{\text{series}}_{b,n} = \phi_{\text{series}}\left(\text{Concat}\left(\mathbf{Z}_{b,n,:}, \mathbf{S}_{n,:}\right)\right). \tag{11}$$

Let $\mathbf{e}^{\text{time}}_{b,h}$ denote the projection of the future temporal covariates and $\mathbf{e}^{\text{step}}_{h}$ the embedding of forecast step h. The three conditions are concatenated for each batch element, forecast step, and series:

$$\mathbf{c}_{b,h,n} = \text{Concat}\left(\mathbf{e}^{\text{time}}_{b,h}, \mathbf{e}^{\text{step}}_{h}, \mathbf{e}^{\text{series}}_{b,n}\right). \tag{12}$$

The fused condition determines a factor-wise scale and shift for the shared future factor representation. The modulation network outputs $2K$ raw values, which are split into raw scale and shift vectors:

$$\left(\boldsymbol{\gamma}^r_{b,h,n}, \boldsymbol{\beta}^r_{b,h,n}\right) = \text{Split}_K\left(g_{\text{mod}}\left(\mathbf{c}_{b,h,n}\right)\right). \tag{13}$$

The bounded modulation parameters and the resulting conditional representation are computed as

$$\begin{aligned} \boldsymbol{\gamma}_{b,h,n} &= \mathbf{1} + 0.1\tanh\left(\boldsymbol{\gamma}^r_{b,h,n}\right), \\ \boldsymbol{\beta}_{b,h,n} &= 0.1\tanh\left(\boldsymbol{\beta}^r_{b,h,n}\right), \\ \widetilde{\mathbf{F}}_{b,h,n,:} &= \boldsymbol{\gamma}_{b,h,n} \odot \widehat{\mathbf{F}}^f_{b,h,:} + \boldsymbol{\beta}_{b,h,n}. \end{aligned} \tag{14}$$

The scale is therefore bounded in $(0.9, 1.1)$ and the shift in $(-0.1, 0.1)$. The same future factor vector can consequently produce different representations across series and forecast steps through the corresponding conditional parameters.

The modulated representation is passed to the load prediction head, which applies normalization and nonlinear projection before producing one scalar forecast for each series and forecast step:

$$\widehat{Y}_{b,h,n} = g_{\text{pred}}\left(\widetilde{\mathbf{F}}_{b,h,n,:}\right). \tag{15}$$

The resulting prediction tensor is $\widehat{\mathbf{Y}} \in \mathbb{R}^{B\times H\times N}$, and the ground-truth tensor is $\mathbf{Y} \in \mathbb{R}^{B\times H\times N}$. Let $\mathbf{M}^f \in \{0,1\}^{B\times H\times N}$ denote the validity mask of future target entries. The forecasting objective is computed in the normalized load space as a masked combination of absolute and squared errors:

$$\mathcal{L} = \frac{\displaystyle\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{n=1}^{N} M^f_{b,h,n}\left[0.7\left|\widehat{Y}_{b,h,n} - Y_{b,h,n}\right| + 0.3\left(\widehat{Y}_{b,h,n} - Y_{b,h,n}\right)^2\right]}{\max\left(1, \displaystyle\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{n=1}^{N} M^f_{b,h,n}\right)}. \tag{16}$$

The historical mask $\mathbf{M}^h$ records the validity of historical inputs, whereas $\mathbf{M}^f$ determines the future target entries included in the forecasting objective.

4 EXPERIMENTS

4.1 DATASETS AND DATA PREPROCESSING

We evaluate STGFM on the Real-world Load Dataset and the UCL Electricity Dataset. Both datasets are organized as aligned load panels with hourly observations, where rows correspond to timestamps and columns correspond to individual load series.

Real-world Load Dataset. The dataset contains hourly consumption measurements from $1{,}257$ industrial load series over $4{,}080$ timestamps, forming an aligned panel of size $4{,}080 \times 1{,}257$.

UCL Electricity Dataset. The dataset contains hourly electricity consumption measurements from multiple metered consumers. The temporal index spans from January 1, 2011 to January 1, 2015, including both endpoints, resulting in $35{,}065$ timestamps. The original quarter-hourly records are aggregated to hourly means. Series are ranked by temporal coverage, with mean absolute activity used as the tie-breaking criterion, and the top 64 series are retained, forming an aligned panel of size $35{,}065 \times 64$.

Data Preprocessing. Raw records are resampled to an hourly grid and assembled into an aligned panel. Missing values are imputed using series-specific means computed from the training split, while missingness information is retained through an observation mask. Each series is standardized using statistics computed from the training split only. Temporal covariates comprise sine and cosine encodings of hour-of-day, day-of-week, and month, together with a weekend indicator. The temporal sequence is partitioned chronologically into training, validation, and test sets with a $70\%/15\%/15\%$ ratio.

4.2 Experimental Setup

We formulate the task as direct multi-horizon forecasting with $H \in \{12, 24, 48, 96\}$ hours. For each dataset and forecasting horizon, a dedicated model predicts all H future steps in a single forward pass. The historical context lengths selected by validation are 96, 240, 240, and 432 hours for the four horizons, respectively. Unless otherwise specified, STGFM uses a model dimension of 128, two interaction layers, four attention heads, dropout of 0.15, and $K = 32$ factors. The entmax parameters for the 12-, 24-, 48-, and 96-hour horizons are 1.0, 1.0, 1.45, and 1.95, respectively. The series-specific static vector contains the training mean, training standard deviation, raw mean, and zero-value ratio of each series.

The training objective combines masked MAE and MSE with weights 0.7 and 0.3 in normalized load space. Models are optimized with AdamW using a learning rate of 2×10^{-4}, weight decay of 10^{-4}, gradient clipping at norm 1.0, and a maximum of 120 epochs. The learning rate is halved after two consecutive epochs without improvement in validation RMSE, and training stops after 15 consecutive epochs with an improvement below 10^{-4}. The checkpoint with the lowest validation RMSE is used for testing. Each stochastic experiment is repeated with five random seeds, and the reported values are the mean and standard deviation across these runs.

4.3 Evaluation Metrics

We evaluate forecasting performance using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). All metrics are computed on the original load scale after inverse normalization, using only observed future target entries indicated by the observation mask.

For the valid target set Ω, the evaluation metrics are defined as

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{|\Omega|}\sum_{i\in\Omega}(y_i-\hat{y}_i)^2, \qquad \mathrm{RMSE} = \sqrt{\mathrm{MSE}}, \\ \mathrm{MAE} &= \frac{1}{|\Omega|}\sum_{i\in\Omega}|y_i-\hat{y}_i|, \qquad R^2 = 1 - \frac{\sum_{i\in\Omega}(y_i-\hat{y}_i)^2}{\sum_{i\in\Omega}(y_i-\bar{y}_\Omega)^2}. \end{aligned} \tag{17}$$

4.4 Main Results

We compare Naive, Seasonal Naive, NLinear, DLinear, LSTM, GRU, TCN, Transformer, Informer, iTransformer, PatchTST, TimesNet, and Mamba, covering naive, linear, recurrent, convolutional, attention-based, and state-space models. All methods use the same datasets, horizons, and chronological test protocol.

Table 1: Comparison results on the Real-world Load Dataset across four prediction horizons.

Horizon	Model	MSE	RMSE	MAE	R^2
12h	STGFM	**0.032774±0.001651**	**0.180979±0.004569**	0.051819±0.001544	**0.965569±0.001735**
12h	TimesNet	0.034135±0.000429	0.184754±0.001161	**0.051492±0.000352**	0.964140±0.000451
12h	DLinear	0.037702±0.000385	0.194168±0.000992	0.055142±0.000133	0.960392±0.000404
12h	NLinear	0.038854±0.000592	0.197109±0.001496	0.055278±0.000291	0.959182±0.000621
24h	STGFM	**0.036476±0.001110**	**0.190964±0.002907**	**0.057055±0.000817**	**0.960853±0.001191**
24h	TimesNet	0.041825±0.000662	0.204506±0.001621	0.058555±0.000470	0.955111±0.000711
24h	DLinear	0.044449±0.000461	0.210826±0.001094	0.060602±0.000209	0.952296±0.000495
24h	NLinear	0.046007±0.000354	0.214490±0.000824	0.061398±0.000368	0.950624±0.000380
48h	STGFM	**0.044270±0.002708**	**0.210308±0.006393**	0.063936±0.001645	**0.952863±0.002883**
48h	TimesNet	0.051181±0.001427	0.226209±0.003181	0.064408±0.000714	0.945505±0.001520
48h	DLinear	0.053427±0.000795	0.231136±0.001721	0.066081±0.000546	0.943114±0.000847
48h	Seasonal Naive	0.057071±0.000000	0.238896±0.000000	**0.062939±0.000000**	0.939233±0.000000
96h	STGFM	**0.051142±0.002535**	**0.226078±0.005590**	0.070370±0.002261	**0.945809±0.002686**
96h	TimesNet	0.057162±0.000921	0.239078±0.001934	**0.068589±0.000450**	0.939431±0.000976
96h	DLinear	0.058884±0.001339	0.242644±0.002766	0.069914±0.000756	0.937606±0.001419
96h	NLinear	0.062198±0.001497	0.249377±0.002988	0.071963±0.001050	0.934095±0.001586

Table 2: Comparison results on the UCL Electricity Dataset across four prediction horizons.

Horizon	Model	MSE ($\times 10^5$)	RMSE	MAE	R^2
12h	STGFM	5.294 ± 0.943	725.39 ± 63.77	86.69 ± 5.39	0.9027 ± 0.0173
12h	TimesNet	5.457 ± 0.101	738.70 ± 6.85	91.24 ± 0.77	0.8997 ± 0.0019
12h	DLinear	5.686 ± 0.077	754.04 ± 5.11	92.14 ± 0.47	0.8955 ± 0.0014
12h	NLinear	5.588 ± 0.230	747.40 ± 15.38	91.94 ± 1.95	0.8973 ± 0.0042
12h	Seasonal Naive	$\mathbf{4.028 \pm 0.000}$	$\mathbf{634.67 \pm 0.00}$	$\mathbf{75.05 \pm 0.00}$	$\mathbf{0.9260 \pm 0.0000}$
24h	STGFM	$\mathbf{7.843 \pm 1.497}$	$\mathbf{882.29 \pm 85.35}$	113.07 ± 7.36	$\mathbf{0.9387 \pm 0.0117}$
24h	TimesNet	10.220 ± 0.235	1010.88 ± 11.68	109.05 ± 1.12	0.9201 ± 0.0018
24h	DLinear	10.875 ± 0.136	1042.83 ± 6.51	110.85 ± 0.76	0.9150 ± 0.0011
24h	NLinear	10.758 ± 0.066	1037.22 ± 3.20	109.82 ± 0.71	0.9159 ± 0.0005
24h	Seasonal Naive	9.231 ± 0.000	960.79 ± 0.00	$\mathbf{101.17 \pm 0.00}$	0.9279 ± 0.0000
48h	STGFM	$\mathbf{8.841 \pm 2.607}$	$\mathbf{932.72 \pm 133.00}$	$\mathbf{107.23 \pm 8.75}$	$\mathbf{0.9309 \pm 0.0204}$
48h	TimesNet	15.423 ± 0.119	1241.89 ± 4.78	125.41 ± 0.43	0.8795 ± 0.0009
48h	DLinear	16.999 ± 0.228	1303.76 ± 8.73	129.72 ± 0.43	0.8671 ± 0.0018
48h	NLinear	17.162 ± 0.134	1310.03 ± 5.10	130.51 ± 0.64	0.8659 ± 0.0010
48h	Seasonal Naive	16.101 ± 0.000	1268.89 ± 0.00	122.47 ± 0.00	0.8742 ± 0.0000
96h	STGFM	$\mathbf{12.240 \pm 1.513}$	$\mathbf{1104.63 \pm 69.20}$	$\mathbf{119.75 \pm 9.66}$	$\mathbf{0.9043 \pm 0.0118}$
96h	TimesNet	22.913 ± 0.248	1513.68 ± 8.20	140.60 ± 0.76	0.8209 ± 0.0019
96h	DLinear	24.520 ± 0.092	1565.89 ± 2.93	144.72 ± 0.70	0.8084 ± 0.0007
96h	NLinear	24.555 ± 0.429	1566.94 ± 13.71	148.00 ± 3.46	0.8081 ± 0.0034
96h	Seasonal Naive	26.324 ± 0.000	1622.45 ± 0.00	151.97 ± 0.00	0.7943 ± 0.0000

Table 1 reports results on the Real-world Load Dataset at prediction horizons of 12, 24, 48, and 96 hours. It presents STGFM and representative baselines, with results for all compared methods provided in the appendix.

Across the four forecasting horizons, STGFM reduces MSE and RMSE and increases R^2 relative to the baselines reported in Table 1. Relative to TimesNet, RMSE decreases by 2.0%, 6.6%, 7.0%, and 5.4% at 12, 24, 48, and 96 hours, respectively. STGFM yields a lower MAE at 24 hours, while TimesNet yields a lower MAE at 12 and 96 hours and Seasonal Naive at 48 hours.

On the UCL Electricity Dataset, Table 2 reports results across four forecasting horizons using the same selected baselines as the Real-world Load Dataset. At 12 hours, STGFM yields lower errors and higher R^2 than the learned baselines, while Seasonal Naive yields better results on all four metrics. At 24 hours, STGFM reduces MSE and RMSE relative to Seasonal Naive by 15.0% and 8.2%, and attains a higher R^2, while its MAE remains higher than those of Seasonal Naive and TimesNet. At 48 and 96 hours, STGFM yields lower errors and higher R^2 than each shared baseline, with RMSE reductions of 21.3% and 16.4% relative to PatchTST.

Table 3: Ablation results across four forecasting horizons on the Real-world Load Dataset and the UCL Electricity Dataset.

Real-world Load Dataset			UCL Electricity Dataset		
Configuration	RMSE	MAE	Configuration	RMSE	MAE
12 h			*12 h*		
STGFM	0.180979 ± 0.004569	0.051819 ± 0.001544	STGFM	725.4 ± 63.8	86.7 ± 5.4
w/o CSE	0.18083 ± 0.00269	0.05256 ± 0.00087	w/o CSE	693.9 ± 59.4	$\mathbf{83.2 \pm 6.1}$
w/Linear	0.18101 ± 0.00547	0.05248 ± 0.00191	w/Linear	737.7 ± 56.5	86.2 ± 6.2
w/LSTM	$\mathbf{0.17957 \pm 0.00373}$	$\mathbf{0.05169 \pm 0.00093}$	w/LSTM	$\mathbf{677.0 \pm 77.8}$	87.5 ± 4.2
w/MLP	0.23376 ± 0.00392	0.07838 ± 0.00132	w/MLP	823.2 ± 157.6	116.4 ± 23.4
w/o Mamba	0.18364 ± 0.00598	0.05257 ± 0.00208	w/o Mamba	718.2 ± 72.9	91.8 ± 9.5
24 h			*24 h*		
STGFM	0.190964 ± 0.002907	$\mathbf{0.057055 \pm 0.000817}$	STGFM	882.3 ± 85.4	113.1 ± 7.4
w/o CSE	0.19887 ± 0.00093	0.05892 ± 0.00047	w/o CSE	1060.8 ± 166.7	120.2 ± 11.8
w/Linear	$\mathbf{0.19079 \pm 0.00139}$	0.05768 ± 0.00162	w/Linear	1163.4 ± 153.9	119.0 ± 13.2
w/LSTM	0.19505 ± 0.00370	0.05870 ± 0.00110	w/LSTM	829.6 ± 86.8	$\mathbf{105.2 \pm 7.3}$
w/MLP	0.23246 ± 0.00422	0.07949 ± 0.00069	w/MLP	1484.8 ± 369.2	191.1 ± 47.6
w/o Mamba	0.19344 ± 0.00704	0.05814 ± 0.00224	w/o Mamba	$\mathbf{824.2 \pm 128.6}$	110.4 ± 12.9
48 h			*48 h*		
STGFM	0.210308 ± 0.006393	0.063936 ± 0.001645	STGFM	$\mathbf{932.7 \pm 133.0}$	107.2 ± 8.7
w/o CSE	0.21623 ± 0.00333	0.06425 ± 0.00148	w/o CSE	940.1 ± 48.3	109.0 ± 7.6
w/Linear	0.21480 ± 0.00696	0.06552 ± 0.00213	w/Linear	1206.5 ± 66.9	115.1 ± 3.2
w/LSTM	$\mathbf{0.20708 \pm 0.00465}$	$\mathbf{0.06271 \pm 0.00141}$	w/LSTM	968.4 ± 144.8	$\mathbf{103.6 \pm 9.5}$
w/MLP	0.22336 ± 0.00710	0.07783 ± 0.00085	w/MLP	1233.8 ± 288.0	142.7 ± 29.9
w/o Mamba	0.20752 ± 0.00139	0.06296 ± 0.00047	w/o Mamba	952.8 ± 44.5	108.9 ± 3.9
96 h			*96 h*		
STGFM	0.226078 ± 0.005590	$\mathbf{0.070370 \pm 0.002261}$	STGFM	1104.6 ± 69.2	119.8 ± 9.7
w/o CSE	0.23195 ± 0.00438	0.07149 ± 0.00156	w/o CSE	1079.3 ± 127.4	125.8 ± 7.0
w/Linear	0.23478 ± 0.00911	0.07385 ± 0.00307	w/Linear	1308.7 ± 189.0	139.3 ± 17.2
w/LSTM	$\mathbf{0.22548 \pm 0.00748}$	0.07079 ± 0.00194	w/LSTM	$\mathbf{981.4 \pm 90.5}$	$\mathbf{114.8 \pm 7.5}$
w/MLP	0.23040 ± 0.01140	0.08233 ± 0.00513	w/MLP	1193.4 ± 127.1	128.7 ± 7.9
w/o Mamba	0.22920 ± 0.01488	0.07172 ± 0.00517	w/o Mamba	1241.3 ± 92.2	128.4 ± 7.6

4.5 Ablation Study

Table 3 evaluates five variants. The `w/o CSE` removes the cross-series encoder, `w/Linear` replaces input-adaptive factor extraction with a linear projection, `w/LSTM` replaces Mamba with an LSTM, `w/MLP` replaces the series-adaptive decoder with a shared MLP, and `w/o Mamba` replaces the dynamics module with a direct factor projection. All other components remain unchanged.

On the Real-world Load Dataset, removing the cross-series encoder increases both RMSE and MAE at every horizon. The linear factor variant increases MAE at all horizons and RMSE at three, with a small RMSE reduction at 24 hours. The shared MLP produces the largest MAE increase, ranging from 17.0% to 51.3%, supporting the role of conditional reconstruction in series- and horizon-specific mapping. On the UCL Electricity Dataset, component effects are horizon dependent. The linear factor variant degrades from 24 to 96 hours, whereas the cross-series encoder and dynamics variants show mixed effects. Removing Mamba increases RMSE from 932.7 to 952.8 at 48 hours and from 1104.6 to 1241.3 at 96 hours, while the shared MLP yields higher MAE at every horizon. Series-adaptive reconstruction therefore shows the most consistent contribution, with factor construction and dynamics providing complementary, horizon-dependent effects.

5 Conclusion

This paper presented STGFM for multi-series, multi-horizon electricity load forecasting. STGFM constructs input-adaptive sparse factors, models their dynamics in a compact space, and reconstructs series-specific forecasts using future temporal information and series context. Across the Real-world Load Dataset and the UCL Electricity Dataset, STGFM improves MSE, RMSE, and R^2 over the evaluated learned baselines, with clearer gains at medium and long horizons on the UCL Electricity Dataset. Ablations identify series-adaptive reconstruction as the most consistent component effect, while cross-series encoding and factor dynamics remain horizon dependent. Future work will consider probabilistic forecasting on larger and more incomplete load panels.

AI USE STATEMENT

Generative AI tools were used solely for language editing and grammar checking. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Ahmad Ahmad, Xun Xiao, Huadong Mo, and Daoyi Dong. TFTformer: A novel transformer based model for short-term load forecasting. *International Journal of Electrical Power & Energy Systems*, 166:110549, 2025.

William Gomez, Fu-Kwun Wang, and Shey-Huei Sheu. Short-term smart grid energy forecasting using a hybrid deep learning method on univariate and multivariate data sets. *Energy*, 335:138081, 2025.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. SOFTS: Efficient multivariate time series forecasting with series-core fusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 64145–64175, 2024.

Yaxuan Kong, Zepu Wang, Yuqi Nie, Tian Zhou, Stefan Zohren, Yuxuan Liang, Peng Sun, and Qingsong Wen. Unlocking the power of LSTM for long term time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11968–11976, 2025.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Tianhao Liu, Fangning Li, Dongdong Zhang, Linke Shan, Hongyu Zhu, Pengcheng Du, Meihui Jiang, and Hui Hwang Goh. Intelligent load forecasting technologies for diverse scenarios in the new power systems. *Renewable and Sustainable Energy Reviews*, 226:116366, 2026.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024. Spotlight.

Yuchen Luo, Xinyu Li, Liuhua Peng, and Mingming Gong. Adapformer: Adaptive channel management for multivariate time series forecasting. *Neural Networks*, 193:107988, 2026.

Lucas Barros Scianni Morais, Giancarlo Aquila, Victor Augusto Durães de Faria, Luana Medeiros Marangon Lima, and José Wanderley Marangon Lima. Short-term load forecasting using neural networks and global climate models: An application to a large-scale electrical power system. *Applied Energy*, 348:121439, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Yunbo Niu, Pei Yong, Juan Yu, and Zhifang Yang. Multivariate power load forecasting model considering meteorological feature-load dynamic forward lag and turning points. *IEEE Transactions on Smart Grid*, 17(3):2490–2505, 2026.

Vasileios Pentsos, Spyros Tragoudas, Jason Wibbenmeyer, and Nasser Khdeer. A hybrid LSTM-Transformer model for power load forecasting. *IEEE Transactions on Smart Grid*, 16(3):2624–2634, 2025.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1185–1196. Association for Computing Machinery, 2025.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

Rajat Sen, Hsiang-Fu Yu, and Inderjit S. Dhillon. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qin, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6607–6617. PMLR, 2019.

Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Xiaocui Yang, Han Zhao, Daling Wang, and Yifei Zhang. Is Mamba effective for time series forecasting? *Neurocomputing*, 619:129178, 2025.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023.

Senzhen Wu, Zhijin Wang, Xiufeng Liu, Yuan Zhao, Yue Hu, and Yaohui Huang. Temporal structure-preserving transformer for industrial load forecasting. *Neural Networks*, 192:107887, 2025.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 753–763. Association for Computing Machinery, 2020.

Guoqi Yu, Jing Zou, Xiaowei Hu, Angelica I. Aviles-Rivero, Jing Qin, and Shujun Wang. Revitalizing multivariate time series forecasting: Learnable decomposition with inter-series dependencies and intra-series variations modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57818–57841, 2024.

Wenhao Zhang, Chun Zhang, Wei Bai, Ning Zhang, Changxia Gao, Yuxin Jia, Chenhao Shi, and Shaoxiong Pang. Unifying channel independence and mixing: Multi-scale patch recursion for global–local representation synergy in multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28427–28436, 2026.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

Yating Zhou and Meng Wang. Empower pre-trained large language models for building-level load forecasting. *IEEE Transactions on Power Systems*, 40(5):4220–4232, 2025.

A Additional Experimental Results

Due to the limited space of the main manuscript, only representative experimental results are presented in the main paper. This appendix provides additional experimental results to further evaluate the effectiveness and robustness of the proposed STGFM model. Specifically, we report the complete comparison results on the private and public datasets, including all baseline models and forecasting horizons.

A.1 Complete Results on the Real-world Load Dataset

In the main manuscript, only partial comparison results on the private dataset are reported to highlight the performance advantages of STGFM. This section presents the complete experimental results involving all compared methods under different forecasting horizons.

Table 4 presents the complete forecasting results on the Real-world Load Dataset across the 12-h, 24-h, 48-h, and 96-h prediction horizons.

The complete results reveal a consistent advantage of STGFM in modeling larger forecasting deviations. Relative to TimesNet, STGFM reduces MSE from 0.034135 to 0.032774, from 0.041825 to 0.036476, from 0.051181 to 0.044270, and from 0.057162 to 0.051142 at 12, 24, 48, and 96 h, corresponding to reductions of 4.0%, 12.8%, 13.5%, and 10.5%, respectively. The corresponding R^2 values increase from 0.964140, 0.955111, 0.945505, and 0.939431 for TimesNet to 0.965569, 0.960853, 0.952863, and 0.945809 for STGFM. MAE is less uniform, with TimesNet or Seasonal Naive yielding smaller values at 12, 48, and 96 h. This metric-specific pattern is consistent with the design of STGFM, which first extracts input-conditioned sparse co-movements as time-indexed factors and models their dynamics in a compact space, then reconstructs each series through series- and forecast-step-conditioned modulation. The resulting factor representation improves the modeling of shared load variations and large deviations, while the conditional decoder preserves the heterogeneous responses of individual series across the forecasting horizon.

A.2 Complete Results on the UCL Electricity Dataset

Table 5 reports the complete comparison on the UCL Electricity Dataset across the 12-, 24-, 48-, and 96-hour forecasting horizons. The table includes STGFM and all baseline models evaluated under the same forecasting protocol. MSE values are scaled by 10^5 for presentation.

The results demonstrate a clear horizon dependent performance pattern. At the 12 hour forecasting horizon, Seasonal Naive achieves lower prediction errors than STGFM, while STGFM consistently outperforms TimesNet among the learning based baselines, reducing RMSE from 738.696 to 725.386 and MAE from 91.240 to 86.686. At the 24 hour forecasting horizon, STGFM substantially improves upon Seasonal Naive in terms of MSE, RMSE, and R^2, reducing MSE from 9.231 to 7.843 and RMSE from 960.787 to 882.291, while increasing R^2 from 0.9279 to 0.9387, although its MAE remains higher than that of Seasonal Naive.

At the 48 hour and 96 hour forecasting horizons, STGFM achieves lower MSE, RMSE, and MAE, together with higher R^2, than all competing methods. Compared with PatchTST, STGFM reduces RMSE by 21.3% and 16.4%, and reduces MSE by 37.1% and 29.9%, at the 48 hour and 96 hour horizons, respectively. This horizon dependent performance pattern is consistent with the underlying design of STGFM, which integrates input adaptive sparse factor construction, compact factor dynamics modeling, and forecast step conditioned series reconstruction. The increasing effectiveness of these components at longer forecasting horizons suggests that the proposed decomposition effectively captures dynamic cross series dependencies that are not adequately represented by fixed seasonal repetition. As the predictive capability of fixed seasonal patterns gradually diminishes with increasing forecasting horizons, the adaptive factor based representation and conditional reconstruction mechanism of STGFM become increasingly important, thereby enabling more effective modeling of long horizon temporal dynamics.

A.3 Additional Analysis

Figures 2 and 3 present qualitative multi-horizon forecasts from STGFM and four representative baselines, namely TimesNet, DLinear, Seasonal Naive, and Transformer. The examples use test series 173655 from the Real-world Load Dataset and test series MT_278 from the UCL Electricity Dataset.

For series 173655, STGFM captures the principal decline, rebound, and evening peak at the shorter horizons. At 48 and 96 hours, it preserves the phase of successive daily cycles and the transitions between low and high load levels, although several peak amplitudes are underestimated. TimesNet and DLinear attenuate the peak responses, Seasonal Naive reproduces a fixed daily profile, and Transformer exhibits larger deviations in timing and amplitude.

Table 4: Complete forecasting results on the Real-world Load Dataset across four prediction horizons.

Horizon	Model	MSE	RMSE	MAE	R^2
12 h	**STGFM**	**0.032774 ± 0.001651**	**0.180979 ± 0.004569**	0.051819 ± 0.001544	**0.965569 ± 0.001735**
	Naive	0.106930 ± 0.000000	0.327002 ± 0.000000	0.098528 ± 0.000000	0.887666 ± 0.000000
	Seasonal Naive	0.043640 ± 0.000000	0.208902 ± 0.000000	0.054282 ± 0.000000	0.954155 ± 0.000000
	NLinear	0.038854 ± 0.000592	0.197109 ± 0.001496	0.055278 ± 0.000291	0.959182 ± 0.000621
	DLinear	0.037702 ± 0.000385	0.194168 ± 0.000992	0.055142 ± 0.000133	0.960392 ± 0.000404
	LSTM	0.061350 ± 0.018582	0.245248 ± 0.034691	0.073757 ± 0.009485	0.935550 ± 0.019521
	GRU	0.045256 ± 0.002258	0.212668 ± 0.005299	0.063589 ± 0.001261	0.952457 ± 0.002372
	TCN	0.055161 ± 0.005631	0.234567 ± 0.011790	0.072453 ± 0.002956	0.942052 ± 0.005916
	Transformer	0.042610 ± 0.001989	0.206366 ± 0.004781	0.059470 ± 0.000877	0.955237 ± 0.002089
	Informer	0.046035 ± 0.002575	0.214476 ± 0.005905	0.065023 ± 0.001765	0.951639 ± 0.002706
	iTransformer	0.041188 ± 0.001626	0.202910 ± 0.004015	0.059757 ± 0.001112	0.956730 ± 0.001708
	PatchTST	0.061823 ± 0.001613	0.248621 ± 0.003252	0.074440 ± 0.000715	0.935053 ± 0.001695
	TimesNet	0.034135 ± 0.000429	0.184754 ± 0.001161	**0.051492 ± 0.000352**	0.964140 ± 0.000451
	Mamba	0.050612 ± 0.004998	0.224698 ± 0.011075	0.069638 ± 0.002429	0.946831 ± 0.005251
24 h	**STGFM**	**0.036476 ± 0.001110**	**0.190964 ± 0.002907**	**0.057055 ± 0.000817**	**0.960853 ± 0.001191**
	Naive	0.158543 ± 0.000000	0.398175 ± 0.000000	0.114330 ± 0.000000	0.829845 ± 0.000000
	Seasonal Naive	0.046958 ± 0.000000	0.216697 ± 0.000000	0.057162 ± 0.000000	0.949603 ± 0.000000
	NLinear	0.046007 ± 0.000354	0.214490 ± 0.000824	0.061398 ± 0.000368	0.950624 ± 0.000380
	DLinear	0.044449 ± 0.000461	0.210826 ± 0.001094	0.060602 ± 0.000209	0.952296 ± 0.000495
	LSTM	0.072799 ± 0.009707	0.269250 ± 0.017419	0.078080 ± 0.003420	0.921869 ± 0.010418
	GRU	0.064777 ± 0.002359	0.254471 ± 0.004643	0.074040 ± 0.000800	0.930479 ± 0.002532
	TCN	0.073754 ± 0.003638	0.271495 ± 0.006647	0.079415 ± 0.001338	0.920844 ± 0.003905
	Transformer	0.058691 ± 0.003816	0.242137 ± 0.007820	0.071363 ± 0.001351	0.937010 ± 0.004095
	Informer	0.067552 ± 0.014356	0.258508 ± 0.026943	0.078064 ± 0.006122	0.927500 ± 0.015407
	iTransformer	0.084273 ± 0.011656	0.289593 ± 0.020218	0.083238 ± 0.006011	0.909555 ± 0.012510
	PatchTST	0.081671 ± 0.001308	0.285771 ± 0.002292	0.084133 ± 0.000406	0.912348 ± 0.001403
	TimesNet	0.041825 ± 0.000662	0.204506 ± 0.001621	0.058555 ± 0.000470	0.955111 ± 0.000711
	Mamba	0.074689 ± 0.009695	0.272714 ± 0.017770	0.081015 ± 0.004771	0.919841 ± 0.010405
48 h	**STGFM**	**0.044270 ± 0.002708**	**0.210308 ± 0.006393**	0.063936 ± 0.001645	**0.952863 ± 0.002883**
	Naive	0.161448 ± 0.000000	0.401806 ± 0.000000	0.116878 ± 0.000000	0.828099 ± 0.000000
	Seasonal Naive	0.057071 ± 0.000000	0.238896 ± 0.000000	**0.062939 ± 0.000000**	0.939233 ± 0.000000
	NLinear	0.058977 ± 0.000765	0.242847 ± 0.001580	0.067634 ± 0.000320	0.937204 ± 0.000815
	DLinear	0.053427 ± 0.000795	0.231136 ± 0.001721	0.066081 ± 0.000546	0.943114 ± 0.000847
	LSTM	0.089964 ± 0.003958	0.299868 ± 0.006605	0.085845 ± 0.002639	0.904211 ± 0.004214
	GRU	0.070311 ± 0.003371	0.265085 ± 0.006374	0.077258 ± 0.000945	0.925136 ± 0.003590
	TCN	0.077830 ± 0.004181	0.278882 ± 0.007420	0.083383 ± 0.001276	0.917130 ± 0.004451
	Transformer	0.070110 ± 0.005363	0.264595 ± 0.009994	0.078145 ± 0.001344	0.925350 ± 0.005711
	Informer	0.073786 ± 0.025851	0.268163 ± 0.043295	0.082136 ± 0.012104	0.921437 ± 0.027525
	iTransformer	0.087225 ± 0.007186	0.295101 ± 0.011861	0.085060 ± 0.004617	0.907127 ± 0.007652
	PatchTST	0.088229 ± 0.001154	0.297027 ± 0.001943	0.087756 ± 0.000483	0.906058 ± 0.001229
	TimesNet	0.051181 ± 0.001427	0.226209 ± 0.003181	0.064408 ± 0.000714	0.945505 ± 0.001520
	Mamba	0.076023 ± 0.006278	0.275489 ± 0.011358	0.082186 ± 0.002358	0.919055 ± 0.006684
96 h	**STGFM**	**0.051142 ± 0.002535**	**0.226078 ± 0.005590**	0.070370 ± 0.002261	**0.945809 ± 0.002686**
	Naive	0.162729 ± 0.000000	0.403397 ± 0.000000	0.119244 ± 0.000000	0.827571 ± 0.000000
	Seasonal Naive	0.067242 ± 0.000000	0.259311 ± 0.000000	0.068970 ± 0.000000	0.928749 ± 0.000000
	NLinear	0.062198 ± 0.001497	0.249377 ± 0.002988	0.071963 ± 0.001050	0.934095 ± 0.001586
	DLinear	0.058884 ± 0.001339	0.242644 ± 0.002766	0.069914 ± 0.000756	0.937606 ± 0.001419
	LSTM	0.093328 ± 0.003638	0.305437 ± 0.005995	0.091682 ± 0.001974	0.901109 ± 0.003855
	GRU	0.084886 ± 0.010201	0.290827 ± 0.017482	0.085176 ± 0.004192	0.910054 ± 0.010809
	TCN	0.085720 ± 0.003513	0.292718 ± 0.006032	0.087723 ± 0.001074	0.909170 ± 0.003723
	Transformer	0.085955 ± 0.006040	0.292999 ± 0.010307	0.085496 ± 0.002403	0.908921 ± 0.006400
	Informer	0.094484 ± 0.037785	0.301603 ± 0.059327	0.094207 ± 0.017332	0.899883 ± 0.040037
	iTransformer	0.097358 ± 0.009714	0.311606 ± 0.016115	0.092038 ± 0.002660	0.896838 ± 0.010294
	PatchTST	0.103261 ± 0.008892	0.321049 ± 0.013748	0.098194 ± 0.006767	0.890583 ± 0.009422
	TimesNet	0.057162 ± 0.000921	0.239078 ± 0.001934	**0.068589 ± 0.000450**	0.939431 ± 0.000976
	Mamba	0.093181 ± 0.012528	0.304559 ± 0.020608	0.092454 ± 0.007290	0.901264 ± 0.013275

For MT_278, STGFM tracks the dominant rising and falling transitions at 12 and 24 hours and retains the daily phase over longer horizons. The competing models show stronger peak attenuation or

Table 5: Complete forecasting results on the UCL Electricity Dataset across four prediction horizons.

Horizon	Model	MSE ($\times 10^5$)	RMSE	MAE	R^2
12 h	STGFM	5.294386 ± 0.942921	725.386477 ± 63.768214	86.685981 ± 5.394805	0.902690 ± 0.017331
	Naive	28.261452 ± 0.000000	1681.114260 ± 0.000000	206.758669 ± 0.000000	0.480560 ± 0.000000
	Seasonal Naive	**4.028008 ± 0.000000**	**634.665932 ± 0.000000**	**75.045406 ± 0.000000**	**0.925966 ± 0.000000**
	NLinear	5.587927 ± 0.230024	747.397822 ± 15.380398	91.935705 ± 1.953989	0.897295 ± 0.004228
	DLinear	5.685916 ± 0.076936	754.036266 ± 5.108153	92.140799 ± 0.471683	0.895494 ± 0.001414
	LSTM	7.631183 ± 0.675905	872.860558 ± 39.254444	111.456151 ± 6.637265	0.859740 ± 0.012423
	GRU	8.128423 ± 0.265466	901.482399 ± 14.655286	118.626200 ± 2.637177	0.850601 ± 0.004879
	TCN	9.480578 ± 0.524622	973.381863 ± 27.054833	134.776035 ± 9.305792	0.825749 ± 0.009642
	Transformer	8.588483 ± 1.649995	923.471834 ± 86.948830	125.824540 ± 9.419703	0.842145 ± 0.030327
	Informer	8.643151 ± 0.604165	929.230377 ± 32.519620	117.877499 ± 3.686967	0.841141 ± 0.011104
	iTransformer	9.128337 ± 1.770276	951.939318 ± 91.140549	118.170554 ± 6.324969	0.832223 ± 0.032537
	PatchTST	7.126226 ± 0.343063	843.976677 ± 20.184945	109.589608 ± 4.066780	0.869021 ± 0.006305
	TimesNet	5.457099 ± 0.101282	738.696372 ± 6.851964	91.239897 ± 0.773081	0.899700 ± 0.001862
	Mamba	7.914081 ± 0.367442	889.417080 ± 20.778321	113.563299 ± 3.291443	0.854541 ± 0.006754
24 h	STGFM	**7.842650 ± 1.496952**	**882.290867 ± 85.351270**	113.066075 ± 7.364831	**0.938706 ± 0.011699**
	Naive	36.913363 ± 0.000000	1921.285062 ± 0.000000	229.669159 ± 0.000000	0.711506 ± 0.000000
	Seasonal Naive	9.231115 ± 0.000000	960.786912 ± 0.000000	**101.165888 ± 0.000000**	0.927855 ± 0.000000
	NLinear	10.758355 ± 0.066343	1037.220965 ± 3.196631	109.819970 ± 0.713754	0.915919 ± 0.000518
	DLinear	10.875363 ± 0.135752	1042.833829 ± 6.513508	110.849712 ± 0.759068	0.915004 ± 0.001061
	LSTM	12.138948 ± 0.549037	1101.545033 ± 24.833897	127.549541 ± 4.055447	0.905129 ± 0.004291
	GRU	11.852382 ± 0.323991	1088.605111 ± 14.877865	126.682993 ± 1.786362	0.907369 ± 0.002532
	TCN	11.800773 ± 0.157178	1086.294423 ± 7.225063	133.484651 ± 2.249326	0.907772 ± 0.001228
	Transformer	12.536129 ± 1.279254	1118.475643 ± 57.284112	139.197733 ± 7.592852	0.902025 ± 0.009998
	Informer	16.284449 ± 2.049062	1273.980223 ± 82.305311	142.942185 ± 12.104461	0.872730 ± 0.016014
	iTransformer	13.274204 ± 1.610727	1150.511743 ± 68.402737	150.866427 ± 9.283613	0.896256 ± 0.012589
	PatchTST	10.911554 ± 0.578426	1044.296895 ± 27.371467	121.964866 ± 4.414297	0.914722 ± 0.004521
	TimesNet	10.219821 ± 0.235266	1010.877295 ± 11.682093	109.052597 ± 1.118520	0.920128 ± 0.001839
	Mamba	13.848884 ± 1.037374	1176.162285 ± 43.742186	135.058969 ± 9.192405	0.891765 ± 0.008108
48 h	STGFM	**8.841133 ± 2.606527**	**932.717776 ± 132.998350**	**107.230098 ± 8.747239**	**0.930903 ± 0.020371**
	Naive	41.102688 ± 0.000000	2027.379791 ± 0.000000	236.259437 ± 0.000000	0.678765 ± 0.000000
	Seasonal Naive	16.100767 ± 0.000000	1268.887990 ± 0.000000	122.467015 ± 0.000000	0.874166 ± 0.000000
	NLinear	17.162078 ± 0.133673	1310.033208 ± 5.098375	130.512334 ± 0.639429	0.865871 ± 0.001045
	DLinear	16.998562 ± 0.227868	1303.761960 ± 8.730787	129.723742 ± 0.430808	0.867149 ± 0.001781
	LSTM	16.195272 ± 0.723145	1272.357225 ± 28.158159	142.639797 ± 4.622574	0.873427 ± 0.005652
	GRU	15.490327 ± 0.238535	1244.571945 ± 9.579861	139.659203 ± 1.174383	0.878937 ± 0.001864
	TCN	15.767277 ± 0.302063	1255.632643 ± 11.958336	149.196958 ± 0.603557	0.876772 ± 0.002361
	Transformer	16.800922 ± 3.203787	1291.751423 ± 119.741858	155.190682 ± 8.498357	0.868694 ± 0.025039
	Informer	18.809156 ± 2.379206	1369.159594 ± 88.865367	157.442975 ± 12.713809	0.852999 ± 0.018594
	iTransformer	16.548217 ± 0.703235	1286.163906 ± 27.479385	166.175508 ± 8.090414	0.870669 ± 0.005496
	PatchTST	14.053704 ± 0.680420	1185.204070 ± 28.759859	134.671831 ± 4.090680	0.890164 ± 0.005318
	TimesNet	15.422975 ± 0.118596	1241.885365 ± 4.779042	125.407778 ± 0.434216	0.879463 ± 0.000927
	Mamba	16.932204 ± 0.725558	1301.003627 ± 27.612596	146.065757 ± 6.979480	0.867668 ± 0.005671
96 h	STGFM	**12.240376 ± 1.513469**	**1104.629847 ± 69.195835**	**119.753320 ± 9.661531**	**0.904336 ± 0.011828**
	Naive	46.838322 ± 0.000000	2164.216294 ± 0.000000	247.022191 ± 0.000000	0.633939 ± 0.000000
	Seasonal Naive	26.323592 ± 0.000000	1622.454683 ± 0.000000	151.974079 ± 0.000000	0.794270 ± 0.000000
	NLinear	24.554578 ± 0.429460	1566.942084 ± 13.705492	148.001248 ± 3.462727	0.808096 ± 0.003356
	DLinear	24.520037 ± 0.091862	1565.885301 ± 2.933355	144.715031 ± 0.703530	0.808366 ± 0.000718
	LSTM	18.820688 ± 2.301015	1369.895099 ± 82.584869	158.139624 ± 13.062361	0.852908 ± 0.017983
	GRU	17.951703 ± 1.184986	1339.238065 ± 44.884405	151.291611 ± 3.322247	0.859700 ± 0.009261
	TCN	19.559861 ± 0.386364	1398.511107 ± 13.818445	163.768962 ± 0.768145	0.847131 ± 0.003020
	Transformer	23.748091 ± 4.397467	1535.593152 ± 144.753138	174.044133 ± 14.619198	0.814399 ± 0.034368
	Informer	21.260593 ± 1.313846	1457.536977 ± 45.349257	167.413927 ± 7.815352	0.833839 ± 0.010268
	iTransformer	19.038749 ± 1.310090	1379.165059 ± 47.152034	164.813792 ± 7.512947	0.851204 ± 0.010239
	PatchTST	17.450777 ± 0.504922	1320.901898 ± 19.230270	152.143223 ± 3.748311	0.863615 ± 0.003946
	TimesNet	22.912922 ± 0.247969	1513.683731 ± 8.196100	140.604793 ± 0.763278	0.820926 ± 0.001938
	Mamba	20.522489 ± 1.586437	1431.740893 ± 54.393955	165.896930 ± 9.188134	0.839608 ± 0.012399

larger deviations near the forecast boundaries. The preservation of turning-point timing and recurrent daily structure is consistent with STGFM's input-adaptive sparse factor construction, compact factor dynamics modeling, and series- and forecast-step-conditioned reconstruction.

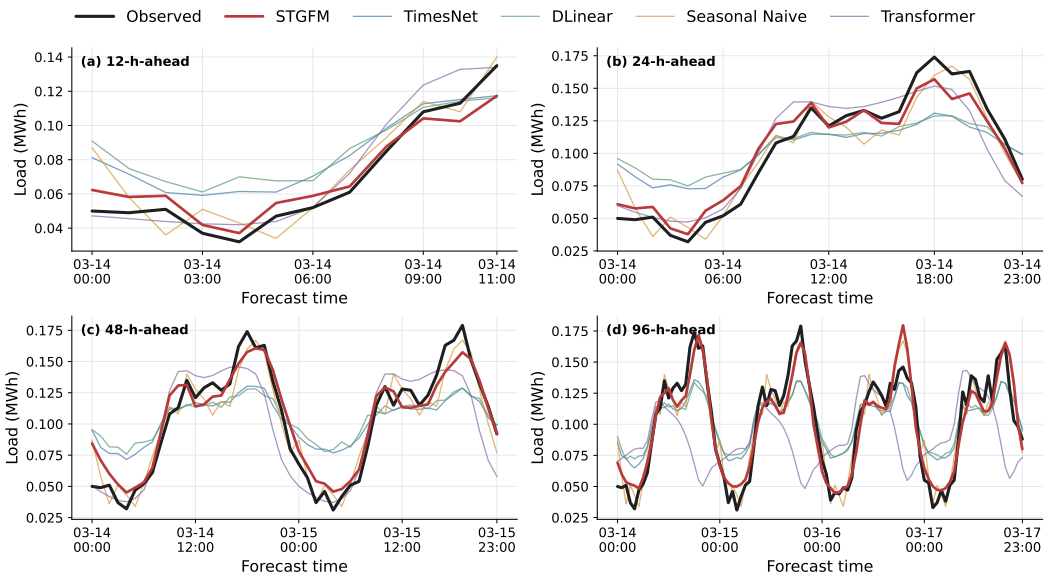


Figure 2: Qualitative multi-horizon forecasts for randomly selected test series 173655 from the Real-world Load Dataset.

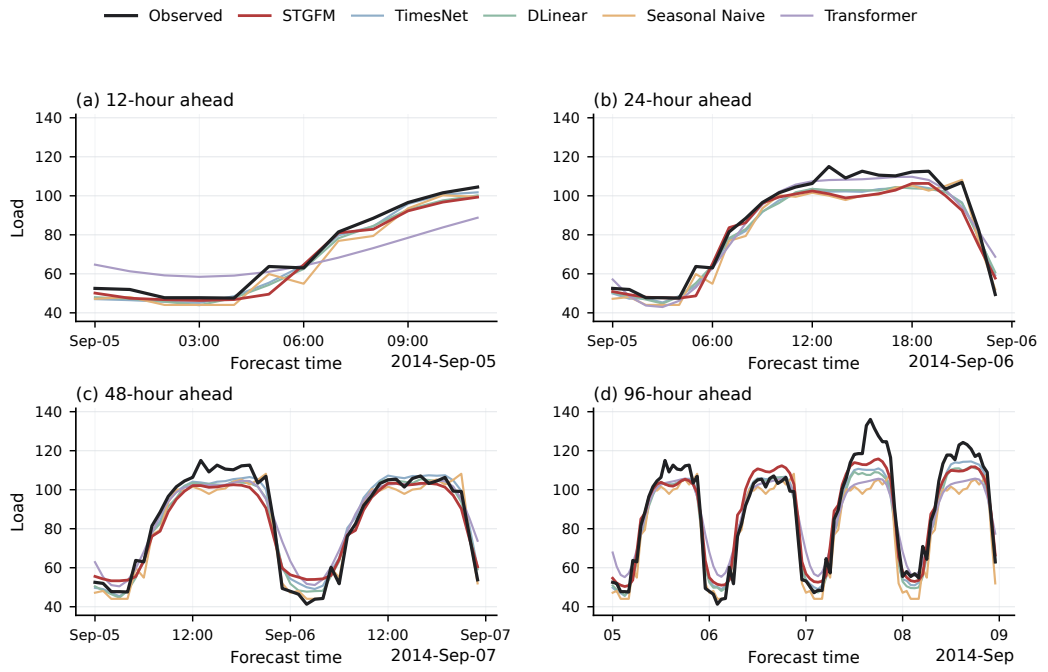


Figure 3: Qualitative multi-horizon forecasts for randomly selected test series MT_278 from the UCL Electricity Dataset.