WxFM-XL: Adapting Univariate Foundation Models to Multi-Station Weather Forecasting

Anonymous authors
Paper under double-blind review

Abstract

With the rise of univariate time series foundation models (*e.g.*, Sundial, Timer), initial efforts have been made to extend them to multivariate settings. However, these models mainly focus on modeling correlations among variables. When they are applied to multi-station weather forecasting, two important factors are often overlooked: (1) the spatial information of stations, and (2) different error priors of different stations relative to the foundation model. In this paper, we propose WxFM-XL, a model for adapting univariate time series foundation models to multi-station weather forecasting. WxFM-XL introduces a cross-station error correlation prior graph to capture stationwise error priors with respect to the foundation model. Building on this, we further propose a dynamic fusion mechanism that adaptively integrates a spatial correlation graph with the error correlation prior graph. Experiments on multiple datasets demonstrate that our model outperforms state of the art baselines.

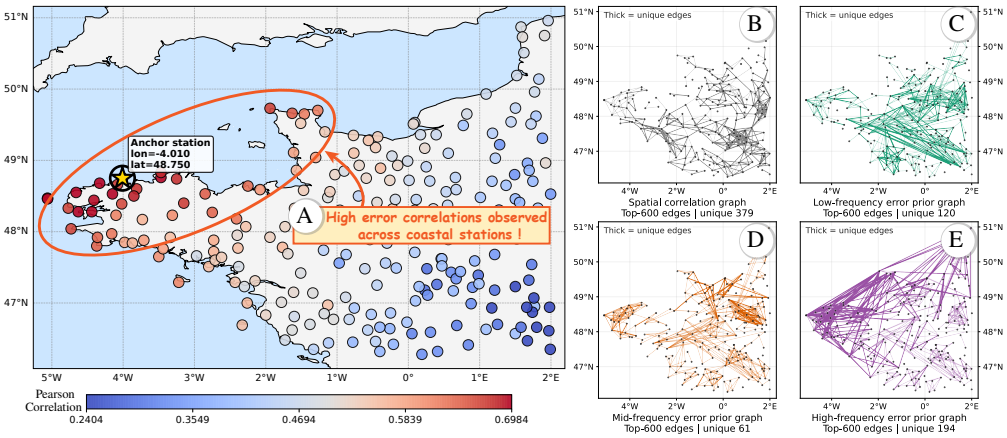
1 Introduction

Building on the recent progress of time series forecasting, time series foundation models have demonstrated strong temporal priors (*e.g.*, Sundial (Liu et al., 2025b), Timer (Liu et al., 2024)). These foundation models learn from large-scale cross-domain data and produce strong univariate forecasts with limited fine tuning, sometimes even in zero-shot settings. However, their performance is limited when directly applied to multi-station weather forecasting. This is because these foundation models are primarily designed for univariate modeling and fall short in capturing the spatiotemporal relationships across stations.

To bridge this gap, some initial efforts have been made to extend univariate foundation models to multivariate settings, such as AdaPTS (Benechehab et al., 2025) and Timer-XL (Liu et al., 2025a). However, directly extending these multivariate methods to multi-station forecasting by treating each station as a variable leads to two limitations. First, these models mainly extend foundation models to modeling correlations among multiple stations but ignore the spatial information of stations (*e.g.*, neighborhood relationships). Existing methods have shown that such spatial information is critical for multi-station forecasting (Wu et al., 2023). Second, stations exhibit different error priors with respect to the foundation model. Recent studies have highlighted that a model's intrinsic errors serve as critical learning signals, implicitly reflecting how well the model captures underlying data structures (Liu et al., 2026). For example, as shown in Fig. 1(a), by calculating the error correlations between a coastal anchor station and all others, we observe complex error correlation priors that extend beyond simple neighborhood relationships. Specifically, the anchor station shares strong error correlations with distant coastal stations (*e.g.*, Fig. 1A). This indicates that relying solely on static station spatial relations is insufficient to capture these foundation model-specific error behaviors.

To address these limitations, rather than treating stations as variables, a method that can natively model the spatial information of stations and their different error priors is desired. However, designing such a method is nontrivial. The spatial information of stations is inherently **static**, remaining constant regardless of the foundation model. In contrast, the error priors are inherently **dynamic**. These error patterns are not fixed. Rather, they are inherently specific to the chosen foundation model and exhibit fluctuations across different frequency bands. The static and dynamic natures

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.



(a) Spatial distribution of Sundial's error correlations between an anchor and all other stations

(b) Edge distributions in Spatial Correlation Graph and Band-Specific Error-Correlation Prior Graph

Figure 1: Visualization of error correlation dependencies for the v-component of wind speed in the French region. (a) Sundial's forecast error correlations relative to an anchor station (star). (b) Topologies of spatial and band-specific error correlation graphs (showing top 600 edges with unique connections highlighted).

of these two types of information make it challenging to jointly model and fuse them in a unified method, since the model must preserve time-invariant spatial dependencies while dynamically adapting to these complex error priors. This inspires us to consider answering the following question in this study: **How can we effectively integrate static station spatial relations with dynamic error priors to improve forecasting performance?**

To address these challenges, we propose WxFM-XL, a model for adapting univariate time series foundation models to multi-station weather forecasting. WxFM-XL first applies frequency domain decomposition to disentangle the foundation model forecasts into distinct frequency components. Subsequently, tailored error correlation prior graphs are constructed for each specific frequency band. As illustrated in Fig. 1(b), these band-specific error priors exhibit distinct topological structures: while the static spatial correlation graph (Fig. 1B) relies solely on geographic proximity, the error correlation prior graphs for low, mid, and high (Figs. 1C- 1E) frequencies reveal nonlocal connections that capture complex patterns beyond simple neighborhood relations (*e.g.*, Fig. 1A). Building on this structural divergence, WxFM-XL introduces a dynamic fusion mechanism that adaptively integrates the spatial correlation graph with these band-specific error correlation prior graphs, resulting in frequency-aware dynamic fusion prior graphs. Guided by these fusion graphs, we implement a lightweight graph residual adapter to extend the foundation model, which effectively injects both the spatial information of stations and the stationwise error priors under a weak fine tuning, while preserving the strong temporal priors of the foundation model. Experiments across multiple datasets demonstrate that our model outperforms the SOTA (state of the art) baselines.

In summary, our contributions are as follows.

- **A foundation model adaptation framework**, which bridges the gap between univariate time series foundation models and multi-station weather forecasting tasks.
- **A dynamic fusion mechanism** that adaptively integrates static station spatial relations with dynamic error priors.
- **Experiments on multiple datasets** demonstrate that our model outperforms SOTA baselines.

2 Related Work

Time series foundation models. Time series foundation models transfer temporal knowledge learned from large-scale, cross-domain corpora to downstream forecasting tasks. For example, Chronos (Ansari et al., 2024) quantizes continuous observations into discrete tokens and formulates

forecasting as language modeling, whereas TimesFM (Das et al., 2024) adopts patch-based modeling to support zero-shot forecasting across diverse datasets. Subsequent studies further improve scalability and generative capability. Time-MoE (Shi et al., 2025) employs sparse mixture-of-experts to scale model capacity without proportionally increasing inference cost, while Timer (Liu et al., 2024) uses a GPT-style architecture and next-token prediction to learn from heterogeneous time series. Sundial (Liu et al., 2025b) further introduces flow matching for probabilistic forecasting over continuous-valued observations, avoiding discrete tokenization. Despite this progress, most foundation models are designed for univariate forecasting, limiting their performance in multi-station weather settings.

Multi-station forecasting models. Multi-station forecasting methods differ mainly in how they represent station interactions and temporal evolution. Graph-based methods represent stations as nodes and learn their spatial interactions. StemGNN (Cao et al., 2020) jointly models inter-series correlations and temporal dynamics in the spectral domain, while MSGNet (Cai et al., 2024) extends this idea with multiscale frequency decomposition and adaptive graph convolution. For long-range temporal modeling, Autoformer (Wu et al., 2021) combines seasonal trend decomposition with an Auto-Correlation mechanism, while FEDformer (Zhou et al., 2022) introduces frequency domain modeling to improve long horizon forecasting. Corrformer (Wu et al., 2023) further combines spatial cross-correlation with temporal autocorrelation for global and regional forecasting. CDPNet (Xu et al., 2025) instead models continuous spatiotemporal evolution from discrete station observations. More recently, STELLA (Fu et al., 2025) replaces Transformer layers with a lightweight MLP to model spatiotemporal correlations without attention. These methods demonstrate the importance of station dependencies, but they generally rely on task-specific architectures trained from scratch and therefore cannot directly exploit the temporal priors acquired through large-scale foundation model pretraining.

Foundation models for multivariate forecasting. Recent studies bridge foundation models and multivariate forecasting through either native multivariate architectures or lightweight adaptation. Moirai (Woo et al., 2024) develops a unified foundation model that supports heterogeneous frequencies and arbitrary numbers of variables. Building on Timer, Timer-XL (Liu et al., 2025a) extends next-token prediction to multivariate series and introduces causal TimeAttention to capture dependencies across variables and long contexts. Rather than modifying the foundation model backbone, AdaPTS (Benechehab et al., 2025) uses lightweight adapters to project multivariate inputs into latent representations that can be processed by existing univariate foundation models. However, when applied to multi-station weather forecasting, these extensions often treat stations as generic variables and lack explicit modeling of the spatial information of stations or the correlation of stationwise error priors relative to the foundation model.

3 Problem Formulation

In this section, we introduce the multi-station weather forecasting problem. For clarity of presentation, we consider forecasting a single weather variable (*e.g.*, temperature). The formulation can be straightforwardly extended to multi-variable forecasting by applying the same procedure to each variable. We study multi-station weather forecasting with N weather stations distributed in a region. Each station has a fixed geographic location (*e.g.*, longitude, latitude, and elevation), and we denote the collection of station locations as $\mathbf{L} \in \mathbb{R}^{N\times 3}$. At each timeframe t, a target weather variable at all stations is denoted by $\mathbf{x}_t \in \mathbb{R}^N$. Given the historical observations over the past T_{in} timeframes $\mathbf{x}_{t-T_{\text{in}}:t-1}$, the goal is to forecast the future weather conditions for the next T_{out} timeframes $\hat{\mathbf{x}}_{t:t+T_{\text{out}}-1} = f(\mathbf{x}_{t-T_{\text{in}}:t-1}, \mathbf{L})$, where $f(\cdot)$ denotes a multi-station forecasting model.

In this work, we focus on adapting a univariate time series foundation model to multi-station weather forecasting. Let $f_{\text{FM}}(\cdot)$ denote a time series foundation model that takes a univariate history and outputs a T_{out}-step forecast for a single station. Applying $f_{\text{FM}}(\cdot)$ independently to N weather stations yields stationwise foundation model forecasts $\tilde{\mathbf{x}}^{\text{FM}}_{t:t+T_{\text{out}}-1} = f_{\text{FM}}(\mathbf{x}_{t-T_{\text{in}}:t-1})$. Our goal is to improve these independent stationwise forecasts by extending the foundation model with a dynamic fusion mechanism and a lightweight graph residual adapter.

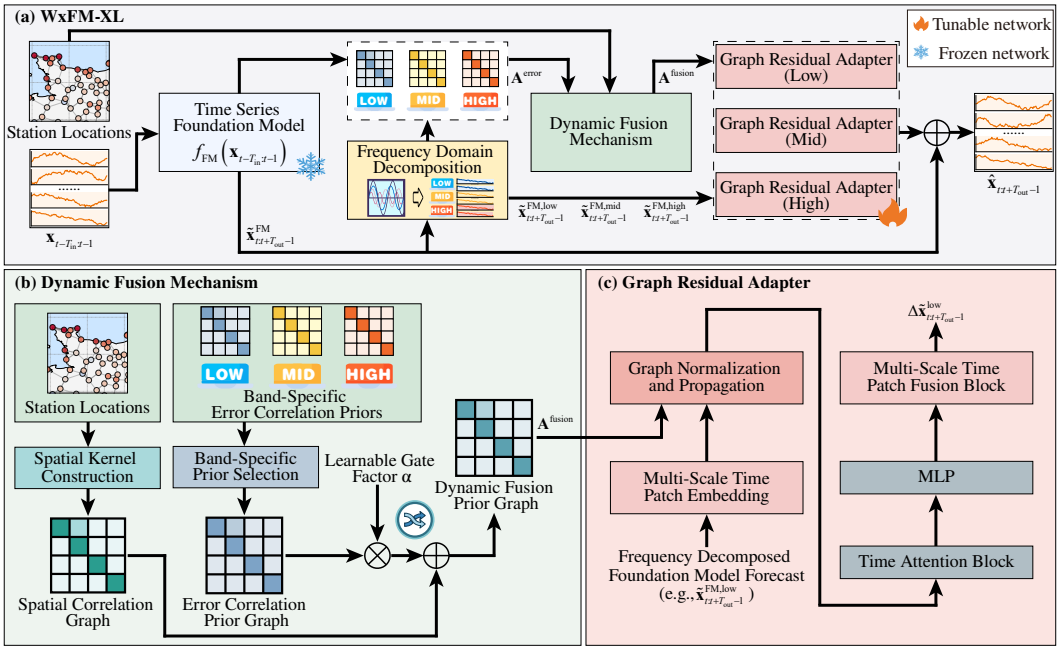


Figure 2: Overview of the WxFM-XL framework.

4 The WxFM-XL Framework

We introduce the WxFM-XL framework for adapting univariate time series foundation models to multi-station weather forecasting. As shown in Fig. 2, WxFM-XL consists of two key components: (1) a **Dynamic Fusion Mechanism** that constructs and combines the spatial correlation graph with error correlation prior graphs; and (2) a **Graph Residual Adapter** that refines foundation model forecasts using the resulting dynamic fusion prior graph while preserving strong temporal priors.

4.1 Dynamic Fusion Mechanism

As shown in Fig. 2(a), the Dynamic Fusion Mechanism first applies **(1) frequency domain decomposition** to decompose the foundation model forecasts into distinct frequency bands (*i.e.*, low, mid, and high frequencies). Then, we construct different **(2) error correlation prior graphs** for each specific frequency band, and the **(3) spatial correlation graph** based on the spatial information. Finally, we develop a gated fusion unit to obtain the **(4) dynamic fusion prior graph**.

(1) Frequency domain decomposition. Foundation model errors exhibit different patterns across frequency bands. To isolate these patterns, we follow frequency domain modeling (Yang et al., 2024) and decompose the foundation model forecasts into low, mid, and high frequency components. Let $\tilde{\mathbf{x}}^{\mathrm{FM}}_{t:t+T_{\mathrm{out}}-1} \in \mathbb{R}^{T_{\mathrm{out}} \times N}$ denote the foundation model forecasts and $\mathcal{F} = \{\mathrm{low}, \mathrm{mid}, \mathrm{high}\}$ the set of frequency bands. For each $f \in \mathcal{F}$, we retain the Fourier coefficients associated with the corresponding frequency band and reconstruct the component as

$$\tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T_{\mathrm{out}}-1} = \mathrm{IDFT}\big(\mathcal{P}_f\big(\mathrm{DFT}\big(\tilde{\mathbf{x}}^{\mathrm{FM}}_{t:t+T_{\mathrm{out}}-1}\big)\big)\big), \qquad f \in \mathcal{F}, \tag{1}$$

where $\mathrm{DFT}(\cdot)$ transforms the forecasts into the frequency domain, $\mathcal{P}_f$ retains the Fourier coefficients assigned to frequency band f and their conjugate counterparts, and $\mathrm{IDFT}(\cdot)$ reconstructs the corresponding component in the time domain. The detailed decomposition and reconstruction are provided in Appendix A.1. These components are subsequently used to construct frequency dependent error correlation prior graphs.

(2) Error correlation prior graph. Stationwise error correlations vary across frequency bands, such that the same station pair may exhibit different dependencies in different frequency components (*e.g.*, Fig. 1(b)). Therefore, for each $f \in \mathcal{F}$, we construct an error correlation prior graph $\mathbf{A}^{\mathrm{error}}_f \in$

$\mathbb{R}^{N\times N}$. We collect the forecast errors associated with frequency band f over the training set and flatten the sample and lead time dimensions into $\mathbf{E}^f \in \mathbb{R}^{M\times N}$, where M is the number of flattened instances and $E^f_{m,i}$ denotes the error at station i for instance m. The adjacency weight between stations i and j is defined by their Pearson correlation:

$$\left(\mathbf{A}_f^{\text{error}}\right)_{i,j} = \frac{\sum_{m=1}^{M}\left(E^f_{m,i}-\bar{E}^f_i\right)\left(E^f_{m,j}-\bar{E}^f_j\right)}{\sqrt{\sum_{m=1}^{M}\left(E^f_{m,i}-\bar{E}^f_i\right)^2}\sqrt{\sum_{m=1}^{M}\left(E^f_{m,j}-\bar{E}^f_j\right)^2}}, \tag{2}$$

where $\bar{E}^f_i = \frac{1}{M}\sum_{m=1}^{M} E^f_{m,i}$ is the mean error at station i. We retain up to k stations with the largest positive Pearson correlations, set non-positive and unselected weights to zero, symmetrize the graph, and set self-loop weights to one.

(3) Spatial correlation graph. We construct a spatial correlation graph from the fixed geographic locations of stations. Let $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_N]^\top \in \mathbb{R}^{N\times 3}$ denote the longitude, latitude, and elevation of the stations. After standardizing the station locations, we construct the spatial adjacency matrix using a Gaussian kernel:

$$\left(\mathbf{A}^{\text{spatial}}\right)_{i,j} = \exp\left(-\frac{\left\|\tilde{\mathbf{l}}_i - \tilde{\mathbf{l}}_j\right\|_2^2}{2\sigma^2}\right), \tag{3}$$

where $\tilde{\mathbf{l}}_i$ is the standardized location of station i, and σ is the median of the nonzero pairwise distances. We retain the k nearest neighbors of each station, including self-loops, set the remaining entries to zero, and symmetrize $\mathbf{A}^{\text{spatial}}$. Since $\mathbf{A}^{\text{spatial}}$ depends only on $\mathbf{L}$, it is static across time.

(4) Dynamic fusion prior graph. As shown in Fig. 2(b), for each frequency band $f \in \mathcal{F}$, we combine the error correlation prior graph $\mathbf{A}_f^{\text{error}}$ with the spatial correlation graph $\mathbf{A}^{\text{spatial}}$ using a learnable gate $\alpha_f \in \mathbb{R}$:

$$\mathbf{A}_f^{\text{fusion}} = \mathbf{A}^{\text{spatial}} + \alpha_f \mathbf{A}_f^{\text{error}}. \tag{4}$$

The spatial graph encodes geographic relations that remain fixed across foundation models, whereas the error graph captures dependencies specific to the selected foundation model and frequency band. By learning α_f for each band, the fusion mechanism balances these static and dynamic priors, preserving geographic constraints while incorporating nonlocal error relations beyond spatial proximity. The resulting graph guides the Graph Residual Adapter in refining $\tilde{\mathbf{x}}^{\text{FM},f}_{t:t+T_{\text{out}}-1}$ according to the error structure of each frequency component.

4.2 Graph Residual Adapter

For each frequency band $f \in \mathcal{F}$, we employ an independent lightweight Graph Residual Adapter to refine $\tilde{\mathbf{x}}^{\text{FM},f}_{t:t+T_{\text{out}}-1}$ under the corresponding dynamic fusion prior graph $\mathbf{A}_f^{\text{fusion}}$, while keeping the foundation model frozen. As shown in Fig. 2(c), we first process the input with the **(1) multi-scale time patch embedding**. Then, we perform the **(2) graph normalization and propagation** to aggregate information on the dynamic fusion prior graph. Finally, the **(3) multi-scale patch fusion block** combines representations from different temporal scales. All these modules are trained in an **(4) end-to-end training** manner.

(1) Multiscale time patch embedding. Since the frozen foundation model already provides strong temporal priors, we use time patching to efficiently extract local residual patterns. For each patch size $p \in \mathcal{P}$ (*e.g.*, $\mathcal{P} = \{6, 12, 24\}$), the frequency component $\tilde{\mathbf{x}}^{\text{FM},f}_{t:t+T_{\text{out}}-1} \in \mathbb{R}^{T_{\text{out}}\times N}$ is divided into nonoverlapping temporal segments and projected into latent tokens:

$$\mathbf{Z}^{(p)} = \text{Reshape}_p\left(\tilde{\mathbf{x}}^{\text{FM},f}_{t:t+T_{\text{out}}-1}\right)\mathbf{W}_p + \mathbf{b}_p. \tag{5}$$

Here, $\text{Reshape}_p(\cdot)$ produces $T_p = T_{\text{out}}/p$ patches with shape $\mathbb{R}^{N\times T_p\times p}$, while $\mathbf{W}_p \in \mathbb{R}^{p\times D}$ and $\mathbf{b}_p \in \mathbb{R}^D$ are learnable parameters. The resulting representation $\mathbf{Z}^{(p)} \in \mathbb{R}^{N\times T_p\times D}$ is passed to the graph propagation module. Using multiple patch sizes enables the adapter to capture residual patterns at different temporal granularities.

(2) Graph normalization and propagation. Given the frequency-specific dynamic fusion prior graph $\mathbf{A}_f^{\text{fusion}}$, we use a learnable node gate $\mathbf{g} \in \mathbb{R}^N$ to adapt its propagation weights. For each patch size $p \in \mathcal{P}$, graph normalization and propagation are defined as

$$\begin{aligned} \widetilde{\mathbf{A}}_f^{\text{fusion}} &= \mathrm{RowNorm}\big(\mathbf{A}_f^{\text{fusion}} \odot \mathrm{sig}\big(\mathbf{g}\mathbf{1}^\top + \mathbf{1}\mathbf{g}^\top\big)\big), \\ \widetilde{\mathbf{Z}}^{(p)} &= \widetilde{\mathbf{A}}_f^{\text{fusion}}\, \mathrm{TokenNorm}\Big(\mathbf{Z}^{(p)}\Big), \qquad p \in \mathcal{P}. \end{aligned} \tag{6}$$

Here, $\mathrm{sig}(\cdot)$ denotes the sigmoid function, and $\mathrm{RowNorm}(\cdot)$ normalizes each row to sum to one. The multiplication is performed along the station dimension, producing $\widetilde{\mathbf{Z}}^{(p)} \in \mathbb{R}^{N \times T_p \times D}$ by aggregating residual information across stations. These graph-propagated representations are then used for temporal refinement and residual reconstruction.

(3) Multi-scale patch fusion block. After graph propagation, we obtain the token representation $\widetilde{\mathbf{Z}}^{(p)}$ for each patch size $p \in \mathcal{P}$. We further refine these representations using a lightweight temporal module consisting of a Time Attention block followed by the MLP. This module captures interactions along the temporal token dimension while keeping the adapter lightweight. The refined representations are mapped back to the temporal domain through the corresponding patch unembedding operation, producing a residual forecast $\Delta\tilde{\mathbf{x}}^{(p,f)}_{t:t+T_{\text{out}}-1} \in \mathbb{R}^{T_{\text{out}} \times N}$ for patch size p and frequency band f. Details of the temporal module, patch unembedding, and scale weight scorer are provided in Appendix A.2.

To combine residuals obtained from different patch sizes, we compute normalized, nonnegative, and sample-dependent weights $\theta^{(p)}$ and perform weighted fusion:

$$\Delta\tilde{\mathbf{x}}^{f}_{t:t+T_{\text{out}}-1} = \sum_{p \in \mathcal{P}} \theta^{(p)} \Delta\tilde{\mathbf{x}}^{(p,f)}_{t:t+T_{\text{out}}-1}. \tag{7}$$

Finally, the adapter combines the residual corrections from all frequency bands. A learnable factor $\beta_f \in \mathbb{R}$ controls the contribution of each band:

$$\hat{\mathbf{x}}_{t:t+T_{\text{out}}-1} = \sum_{f \in \mathcal{F}} \Big(\tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T_{\text{out}}-1} + \beta_f \Delta\tilde{\mathbf{x}}^{f}_{t:t+T_{\text{out}}-1}\Big). \tag{8}$$

(4) End-to-end training. We keep the foundation model frozen and jointly optimize all adapter parameters, including the node gates $\mathbf{g}$, prior fusion factors α_f, and band scaling factors β_f. Because low-frequency components usually contain substantially more energy, a standard forecasting loss may underemphasize errors in high-frequency components. We therefore combine the MSE of the final forecast with an energy-normalized RMSE for each frequency band.

For clarity, we omit the forecast-window subscripts and define $\hat{\mathbf{x}}^f = \tilde{\mathbf{x}}^{\mathrm{FM},f} + \beta_f \Delta\tilde{\mathbf{x}}^f$. The training objective is

$$\mathcal{L} = \lambda_{\text{mse}} \mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) + \sum_{f \in \mathcal{F}} \lambda_f \sqrt{\frac{\|\hat{\mathbf{x}}^f - \mathbf{x}^f\|_{\mathrm{F}}^2}{\|\mathbf{x}^f\|_{\mathrm{F}}^2 + \varepsilon}}. \tag{9}$$

Here, $\mathbf{x}^f$ denotes the ground-truth component obtained using the same frequency decomposition, $\varepsilon > 0$ ensures numerical stability, and λ_{mse} and λ_f balance the two loss terms. Energy normalization prevents low-energy frequency components from being overwhelmed during training, enabling balanced residual correction across frequency bands.

5 Experiments

5.1 Experimental Settings

Datasets. We evaluate all methods on three hourly weather station datasets. The MeteoNet dataset (Larvor et al., 2020), provided by the French national weather service, contains observations from 234 stations between January 1 and December 31, 2016, including the u-component and v-component of wind at 10 meters and temperature at 2 meters. The Hunan dataset, provided by the

Table 1: MSE results on the Hunan regional, French regional, and global datasets with a 96-hour input window and forecast horizons of 24 and 96 hours. Best and second-best results are shown in **bold** and underlined, respectively, based on unrounded values. Complete MAE results are reported in the appendix.

Models			WxFM-XL (Timer)	WxFM-XL (Sundial)	Timer	Sundial	AdaPTS (Timer)	AdaPTS (Sundial)	$\text{Moirai}_{\text{ft}}$	Moirai	$\text{Timer-XL}_{\text{ft}}$	Timer-XL	STELLA	CDPNet	MSGNet	Corrformer
Hunan	temp	24	$\underline{5.14}_{\pm 0.01}$	$\mathbf{5.06}_{\pm 0.09}$	5.28	5.78	$5.33_{\pm 0.05}$	$5.57_{\pm 0.11}$	$5.65_{\pm 0.27}$	5.24	$5.43_{\pm 0.15}$	5.26	$8.56_{\pm 0.84}$	$7.44_{\pm 0.79}$	$5.89_{\pm 0.39}$	$6.01_{\pm 0.47}$
		96	$\mathbf{7.31}_{\pm 0.04}$	$\underline{7.33}_{\pm 0.05}$	7.52	8.35	$7.73_{\pm 0.42}$	$8.51_{\pm 0.51}$	$7.54_{\pm 0.29}$	7.52	$9.29_{\pm 0.27}$	7.37	$16.06_{\pm 1.97}$	$26.94_{\pm 6.19}$	$12.45_{\pm 0.93}$	$8.60_{\pm 0.48}$
	v-wind	24	$\underline{2.45}_{\pm 0.03}$	$\mathbf{2.43}_{\pm 0.01}$	2.63	3.07	$2.52_{\pm 0.03}$	$2.54_{\pm 0.14}$	$2.76_{\pm 0.02}$	2.81	$2.58_{\pm 0.01}$	2.58	$2.50_{\pm 0.03}$	$2.63_{\pm 0.09}$	$2.69_{\pm 0.07}$	$3.46_{\pm 0.12}$
		96	$\mathbf{3.10}_{\pm 0.00}$	$3.30_{\pm 0.08}$	$\underline{3.12}$	3.66	$3.23_{\pm 0.11}$	$3.27_{\pm 0.16}$	$3.62_{\pm 0.32}$	3.83	$3.37_{\pm 0.37}$	3.20	$3.43_{\pm 0.34}$	$4.54_{\pm 1.41}$	$4.62_{\pm 0.28}$	$3.95_{\pm 0.09}$
	u-wind	24	$2.08_{\pm 0.00}$	$\mathbf{2.02}_{\pm 0.00}$	2.16	2.57	$2.14_{\pm 0.13}$	$2.16_{\pm 0.02}$	$2.38_{\pm 0.01}$	2.37	$2.14_{\pm 0.01}$	2.16	$\underline{2.03}_{\pm 0.01}$	$2.15_{\pm 0.03}$	$2.30_{\pm 0.12}$	$2.14_{\pm 0.05}$
		96	$\underline{2.15}_{\pm 0.03}$	$\mathbf{2.15}_{\pm 0.11}$	2.35	2.80	$2.31_{\pm 0.04}$	$2.29_{\pm 0.13}$	$2.71_{\pm 0.03}$	2.82	$2.27_{\pm 0.03}$	2.35	$2.21_{\pm 0.00}$	$2.51_{\pm 0.09}$	$2.89_{\pm 0.15}$	$2.39_{\pm 0.19}$
French	temp	24	$\underline{6.57}_{\pm 0.07}$	$\mathbf{6.50}_{\pm 0.10}$	7.57	8.02	$7.29_{\pm 0.21}$	$7.31_{\pm 0.15}$	$7.23_{\pm 0.05}$	7.51	$6.69_{\pm 0.32}$	7.07	$8.36_{\pm 0.21}$	$22.42_{\pm 3.87}$	$6.98_{\pm 0.35}$	$9.81_{\pm 0.57}$
		96	$\mathbf{12.08}_{\pm 0.07}$	$12.54_{\pm 0.39}$	$\underline{12.35}$	13.52	$12.92_{\pm 0.74}$	$13.52_{\pm 0.43}$	$12.73_{\pm 1.23}$	12.85	$14.17_{\pm 2.14}$	12.52	$19.22_{\pm 2.20}$	$83.78_{\pm 3.77}$	$13.14_{\pm 0.41}$	$13.37_{\pm 0.51}$
	v-wind	24	$\mathbf{4.50}_{\pm 0.03}$	$\underline{4.54}_{\pm 0.01}$	5.69	6.11	$4.87_{\pm 0.21}$	$5.24_{\pm 0.23}$	$5.39_{\pm 0.14}$	6.30	$6.01_{\pm 0.05}$	5.66	$5.28_{\pm 0.03}$	$5.15_{\pm 0.35}$	$5.31_{\pm 0.14}$	$6.06_{\pm 0.39}$
		96	$\mathbf{6.39}_{\pm 0.04}$	$\underline{6.60}_{\pm 0.28}$	7.96	9.08	$6.98_{\pm 0.42}$	$7.18_{\pm 0.37}$	$8.15_{\pm 0.37}$	10.80	$8.46_{\pm 0.29}$	7.91	$7.21_{\pm 0.24}$	$7.83_{\pm 0.88}$	$8.95_{\pm 0.59}$	$8.14_{\pm 0.50}$
	u-wind	24	$\mathbf{4.60}_{\pm 0.03}$	$\underline{4.63}_{\pm 0.01}$	6.26	7.10	$5.33_{\pm 0.21}$	$5.70_{\pm 0.30}$	$5.65_{\pm 0.07}$	6.17	$4.90_{\pm 0.06}$	6.54	$5.79_{\pm 0.16}$	$5.36_{\pm 1.07}$	$5.09_{\pm 0.15}$	$6.08_{\pm 0.40}$
		96	$\underline{8.21}_{\pm 0.13}$	$\mathbf{8.11}_{\pm 0.14}$	9.79	11.45	$8.29_{\pm 0.31}$	$8.60_{\pm 0.39}$	$10.11_{\pm 0.53}$	12.40	$8.81_{\pm 0.34}$	9.94	$8.85_{\pm 0.44}$	$9.28_{\pm 0.19}$	$9.71_{\pm 0.72}$	$9.51_{\pm 0.26}$
Global	temp	24	$\underline{7.57}_{\pm 0.01}$	$\mathbf{7.49}_{\pm 0.03}$	9.67	10.27	$8.73_{\pm 0.52}$	$9.15_{\pm 0.17}$	$9.07_{\pm 0.06}$	9.87	$8.04_{\pm 0.60}$	10.90	$7.61_{\pm 0.01}$	$8.81_{\pm 0.19}$	$9.90_{\pm 0.31}$	$7.79_{\pm 0.46}$
		96	$\mathbf{13.95}_{\pm 0.08}$	$\underline{13.99}_{\pm 0.01}$	16.47	17.94	$15.77_{\pm 0.45}$	$16.15_{\pm 0.17}$	$16.24_{\pm 0.07}$	18.55	$14.36_{\pm 0.14}$	18.11	$14.46_{\pm 0.18}$	$14.93_{\pm 2.44}$	$17.97_{\pm 1.35}$	$16.34_{\pm 0.36}$
	wind	24	$\mathbf{3.58}_{\pm 0.01}$	$\underline{3.66}_{\pm 0.00}$	4.11	4.91	$4.03_{\pm 0.15}$	$4.17_{\pm 0.12}$	$4.24_{\pm 0.02}$	4.83	$3.76_{\pm 0.06}$	4.18	$3.73_{\pm 0.01}$	$4.17_{\pm 0.07}$	$5.25_{\pm 0.39}$	$3.74_{\pm 0.17}$
		96	$\mathbf{4.69}_{\pm 0.10}$	$\underline{4.76}_{\pm 0.02}$	5.26	6.23	$4.81_{\pm 0.22}$	$4.94_{\pm 0.15}$	$5.42_{\pm 0.03}$	7.36	$4.79_{\pm 0.01}$	5.23	$4.85_{\pm 0.02}$	$4.98_{\pm 0.14}$	$6.97_{\pm 0.06}$	$5.14_{\pm 0.18}$

Hunan Province Meteorological Bureau, contains the same three variables from 3,570 stations between April 1 and September 30, 2023. The global dataset, obtained from the National Centers for Environmental Information (NCEI) (NOAA National Centers for Environmental Information, 2026; Wu et al., 2023), contains hourly averaged wind speed and temperature from 3,850 stations between January 1, 2019, and December 31, 2020. For all datasets, we chronologically split the observations into training, validation, and test sets using a 6:1:3 ratio. Detailed dataset information is provided in Appendix B.1.

Baselines. We compare WxFM-XL with three groups of forecasting methods. **(1) Time series foundation models.** Timer (Liu et al., 2024) and Sundial (Liu et al., 2025b) are evaluated under zero-shot settings without parameter updates on the target datasets. **(2) Multi-station forecasting models.** STELLA (Fu et al., 2025), CDPNet (Xu et al., 2025), MSGNet (Cai et al., 2024), and Corrformer (Wu et al., 2023) are included as models that explicitly capture dependencies across stations. **(3) Foundation models for multivariate forecasting.** Timer-XL (Liu et al., 2025a) and Moirai (Woo et al., 2024) are evaluated under both zero-shot and fine-tuning settings, with their fine-tuned variants denoted as $\text{Timer-XL}_{\text{ft}}$ and $\text{Moirai}_{\text{ft}}$. AdaPTS (Benechehab et al., 2025) is combined with Timer and Sundial, yielding AdaPTS (Timer) and AdaPTS (Sundial), respectively. Detailed baseline information is provided in Appendix B.2.

Implementation details. All baselines follow their official implementations and recommended settings. We use frozen Timer and Sundial backbones, denoted as WxFM-XL (Timer) and WxFM-XL (Sundial). To accommodate Timer (Liu et al., 2024), all models use a 96-hour input window, with forecast horizons $T_{\text{out}} \in \{24, 96\}$. We train WxFM-XL with Adam for up to 20 epochs, using a cosine learning rate schedule from 10^{-4} to 10^{-6}.

Stochastic experiments use three seeds (2021, 2022, and 2023); for trainable methods, we report the mean and sample standard deviation. Timer and Timer-XL use deterministic inference, while Sundial and Moirai show negligible variation at the reported precision, so their standard deviations are omitted. WxFM-XL uses three frequency bands $\mathcal{F} = \{\text{low}, \text{mid}, \text{high}\}$ and patch sizes $\mathcal{P} = \{6, 12, 24\}$. Further settings are provided in Appendix B.3.

Table 2: Ablation results (MSE) on the Hunan regional and French regional datasets with a 96-hour input window and a 24-hour forecast horizon. Results are reported as the mean and sample standard deviation over three random seeds. Best results are shown in **bold**. Complete MAE results are provided in the appendix.

Model variant	Hunan regional			French regional		
	temp	v-wind	u-wind	temp	v-wind	u-wind
WxFM-XL (Timer)	**5.143**$_{\pm 0.013}$	**2.450**$_{\pm 0.034}$	**2.084**$_{\pm 0.003}$	**6.571**$_{\pm 0.065}$	**4.501**$_{\pm 0.032}$	**4.602**$_{\pm 0.028}$
w/o frequency domain decomposition	5.393$_{\pm 0.022}$	2.559$_{\pm 0.067}$	2.140$_{\pm 0.011}$	6.773$_{\pm 0.065}$	4.696$_{\pm 0.032}$	4.900$_{\pm 0.024}$
w/o error correlation prior graph	5.430$_{\pm 0.033}$	2.534$_{\pm 0.019}$	2.110$_{\pm 0.002}$	6.808$_{\pm 0.124}$	4.672$_{\pm 0.073}$	4.882$_{\pm 0.025}$
w/o multi-scale temporal module	5.671$_{\pm 0.036}$	2.616$_{\pm 0.035}$	2.149$_{\pm 0.021}$	6.986$_{\pm 0.108}$	5.251$_{\pm 0.056}$	5.551$_{\pm 0.035}$

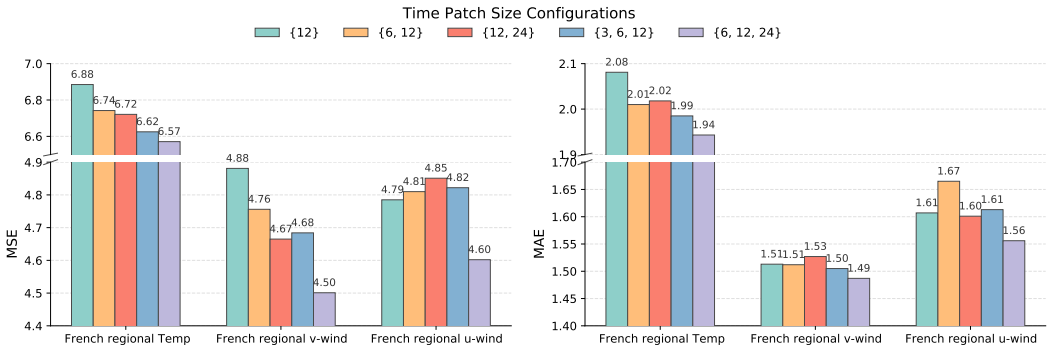


Figure 3: Performance comparison of WxFM-XL (Timer) under different time patch size configurations. The look-back window is set to 96, and the prediction horizon is 24.

5.2 Result Analysis

From the results in Table 1, we draw three observations.

(1) **Overall performance.** Across all 16 dataset-variable-horizon settings, a WxFM-XL variant achieves the best MSE: WxFM-XL (Timer) ranks first in nine settings and WxFM-XL (Sundial) in the remaining seven. Both variants consistently outperform their frozen backbones, demonstrating the effectiveness of WxFM-XL across different regions, variables, and forecasting horizons.

(2) **Comparison with task-specific models.** Compared with STELLA, CDPNet, MSGNet, and Corrformer, WxFM-XL generally achieves better performance. On the four French wind forecasting settings, the best WxFM-XL variant reduces MSE by 8.3%–12.6% relative to the strongest task-specific baseline. These results indicate that combining pretrained temporal priors with spatial relations and frequency-specific error correlations provides more effective cross-station modeling than task-specific architectures trained from scratch.

(3) **Comparison with foundation model extensions.** Fine tuning improves the average performance of both Timer-XL and Moirai, but the gains are not uniform across forecasting tasks. For example, Timer-XL fine tuning increases the French 24-hour v-wind MSE from 5.66 to 6.01 and the Hunan 96-hour temperature MSE from 7.37 to 9.29. This variation indicates that generic multivariate fine tuning does not uniformly adapt to the heterogeneous dependencies among weather stations. AdaPTS also generally underperforms WxFM-XL, indicating that generic multivariate adaptation is less reliable than the proposed station-aware, frequency-specific residual correction.

5.3 Ablation Study

We conduct component ablations on the Hunan and French regional datasets to evaluate the core designs of WxFM-XL. Table 2 compares the full model using Timer with three ablated variants.

Impact of frequency domain decomposition. Removing frequency domain decomposition consistently degrades performance. For example, the MSE values for Hunan temperature and u-wind

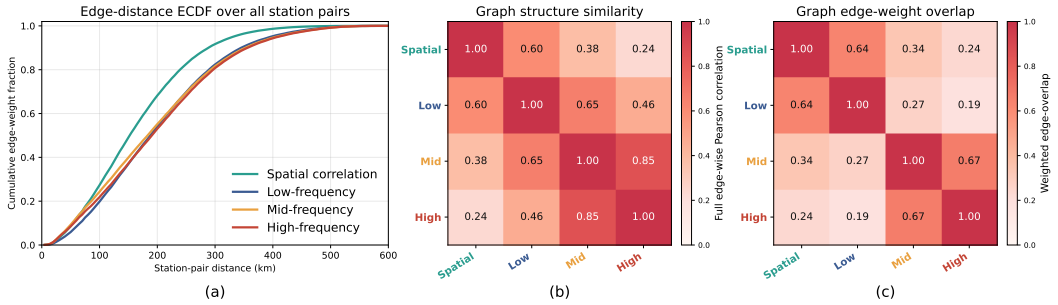


Figure 4: Structural properties of prior graphs (taking the v-component of wind speed in the French region as an example).

increase from 5.143 to 5.393 and from 2.084 to 2.140, respectively. This result shows that separating forecasts into frequency bands reduces cross-band interference and supports band-specific residual correction.

Impact of the error correlation prior graph. Removing the error correlation prior graph leaves only the static spatial graph for message passing. Consequently, the v-wind MSE increases from 2.450 to 2.534 on the Hunan dataset and from 4.501 to 4.672 on the French dataset. This degradation indicates that geographic proximity alone cannot adequately represent the model-specific and nonlocal dependencies among forecast errors.

Impact of the multi-scale temporal module. Removing the multi-scale temporal module increases the Hunan temperature MSE from 5.143 to 5.671 and the French u-wind MSE from 4.602 to 5.551. Figure 3 further shows that the $\{6, 12, 24\}$ configuration achieves the lowest MSE and MAE across all three French regional variables, confirming the benefit of combining complementary temporal resolutions.

5.4 Visualization Study

We visualize the graph structures to examine how the frequency-specific error priors complement the static spatial graph. Fig. 4 compares their edge-distance distributions, structural correlations, and weighted edge overlaps.

Nonlocal dependencies in error priors. Fig. 4(a) shows the edge-distance weighted ECDF over all station pairs. The spatial graph concentrates its edge weights at short geographic distances, whereas the error prior graphs exhibit heavier tails. This difference indicates that foundation model errors contain correlated patterns across geographically distant stations, allowing the error graphs to provide nonlocal connections absent from purely spatial message passing.

Frequency-dependent graph structures. Fig. 4(b) and (c) show that the error prior graphs differ across frequency bands and increasingly diverge from the spatial graph. In particular, their Pearson correlation with the spatial graph decreases from 0.60 at the low-frequency band to 0.24 at the high-frequency band. These results indicate that a single static graph cannot represent error dependencies across all frequency components and support the dynamic fusion of spatial and frequency-specific error graphs.

6 Conclusion

We proposed WxFM-XL to adapt univariate time series foundation models to multi-station weather forecasting. WxFM-XL decomposes foundation model forecasts into frequency bands and dynamically integrates band-specific error correlation priors with static spatial relations. A lightweight graph residual adapter then performs structured cross-station correction while preserving the temporal priors of the foundation model. Experiments on regional and global datasets demonstrate consistent improvements over existing baselines. While this work focuses on interactions among stations, future work will further incorporate dependencies among meteorological variables for joint multivariate and multi-station forecasting.

AI USE STATEMENT

Generative AI was used solely for language polishing and grammar correction. No AI tools were involved in the research process, including methodology design, data analysis, or implementation.

REPRODUCIBILITY STATEMENT

The methodological derivations and architectural details required to reproduce WxFM-XL are provided in Appendix A. Dataset preprocessing, baseline configurations, hyperparameter settings, random seeds, hardware, and evaluation protocols are documented in Appendix B. Complete MAE results, ablation results, and forecasting cases are reported in Appendix C. The Supplementary Material includes the core implementation of the proposed framework. The complete source code, configurations, and data processing scripts will be publicly released upon acceptance.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024, 2024.

Abdelhakim Benechehab, Vasilii Feofanov, Giuseppe Paolo, Albert Thomas, Maurizio Filippone, and Balázs Kégl. AdaPTS: Adapting univariate foundation models to probabilistic multivariate time series forecasting. In *Proceedings of the International Conference on Machine Learning*, pp. 3731–3748, Vancouver, BC, Canada, 2025. PMLR.

Wanlin Cai, Yuxuan Liang, Xianggen Liu, Jianshuai Feng, and Yuankai Wu. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11141–11149, Vancouver, BC, Canada, 2024. AAAI Press.

Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Spectral temporal graph neural network for multivariate time-series forecasting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 17766–17778, Red Hook, NY, USA, 2020. Curran Associates Inc.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the International Conference on Machine Learning*, pp. 10148–10167, Vienna, Austria, 2024. PMLR.

Yisong Fu, Fei Wang, Zezhi Shao, Boyu Diao, Lin Wu, Zhulin An, Chengqing Yu, Yujie Li, and Yongjun Xu. On the integration of spatial-temporal knowledge: A lightweight approach to atmospheric time series forecasting. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 82509–82538. Curran Associates, Inc., 2025.

Gwennaëlle Larvor, Léa Berthomier, Vincent Chabot, Brice Le Pape, Bruno Pradel, and Lior Perez. Meteonet, an open reference weather dataset by meteo france. `https://github.com/meteofrance/meteonet`, 2020. Accessed: 2025-01-12.

Jiexi Liu, Meng Cao, and Songcan Chen. Beyond observations: Reconstruction error-guided irregularly sampled time series representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 23712–23720, Singapore, 2026. AAAI Press.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 32369–32399, Vienna, Austria, 2024. PMLR.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *Proceedings of the International Conference on Learning Representations*, pp. 83982–84006, Singapore, 2025a. OpenReview.net.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the International Conference on Machine Learning*, pp. 39295–39317, Vancouver, BC, Canada, 2025b. PMLR.

NOAA National Centers for Environmental Information. Global hourly - integrated surface database (isd). `https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database`, 2026. Accessed: 2026-01-14.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *Proceedings of the International Conference on Learning Representations*, pp. 34635–34667, Singapore, 2025. OpenReview.net.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 53140–53164, Vienna, Austria, 2024. PMLR.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, Red Hook, NY, USA, 2021. Curran Associates, Inc.

Haixu Wu, Hang Zhou, Mingsheng Long, and Jianmin Wang. Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, 5(6):602–611, 2023.

Chujie Xu, Yuqing Ma, Haoyuan Deng, Yajun Gao, Yudie Wang, Kai Lv, and Xianglong Liu. Continuous diffusive prediction network for multi-station weather prediction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 6714–6722, Montreal, Canada, 2025. ijcai.org.

Runze Yang, Longbing Cao, Jianxun Li, and Jie Yang. Rethinking fourier transform from a basis functions perspective for long-term time series forecasting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pp. 8515–8540, Red Hook, NY, USA, 2024. Curran Associates Inc.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the International Conference on Machine Learning*, pp. 27268–27286, Baltimore, MD, USA, 2022. PMLR.

APPENDIX OVERVIEW

This appendix provides additional methodological derivations, implementation details, and experimental results that complement the main paper. Appendix A presents the complete frequency decomposition procedure, details of the Graph Residual Adapter, the training objective, and a complexity analysis. Appendix B describes the datasets, baselines, implementation settings, and evaluation protocol. Appendix C reports the complete MAE results, MAE ablation results, and representative forecasting cases across the evaluated datasets, variables, and forecasting horizons. The notation follows the definitions introduced in the main paper.

A METHOD DETAILS

This section expands the operations summarized in Sections 4.1 and 4.2. We first describe how the foundation model forecasts are separated into frequency-specific components. We then provide the complete processing flow of the Graph Residual Adapter, followed by the optimization objective and computational complexity.

A.1 FREQUENCY DOMAIN DECOMPOSITION

The errors produced by a univariate foundation model may exhibit different dependencies across temporal frequencies. We therefore decompose its forecasts into low-, mid-, and high-frequency components before constructing the error correlation prior graphs.

Let $T = T_{\text{out}}$ and let

$$\tilde{\mathbf{x}}^{\text{FM}}_{t:t+T-1} = \left[\tilde{\mathbf{x}}^{\text{FM}}[0], \ldots, \tilde{\mathbf{x}}^{\text{FM}}[T-1]\right]^{\top} \in \mathbb{R}^{T \times N} \tag{10}$$

denote the foundation model forecast over N stations. Here, $\tilde{\mathbf{x}}^{\text{FM}}[\tau] \in \mathbb{R}^N$ is the forecast at the τ-th step within the output window.

The discrete Fourier transform (DFT) of the forecast is

$$\mathbf{H}[k] = \sum_{\tau=0}^{T-1} \tilde{\mathbf{x}}^{\text{FM}}[\tau] \exp\left(-j\frac{2\pi k\tau}{T}\right), \qquad k = 0, \ldots, T-1, \tag{11}$$

where $j = \sqrt{-1}$ and $\mathbf{H}[k] \in \mathbb{C}^N$ contains the Fourier coefficients of all stations at frequency index k.

Because the forecast is real valued, its Fourier spectrum satisfies the conjugate symmetry property

$$\mathbf{H}[T-k] = \overline{\mathbf{H}[k]}, \tag{12}$$

where $\overline{\mathbf{H}[k]}$ denotes the complex conjugate of $\mathbf{H}[k]$. It is therefore sufficient to partition the nonnegative frequency indices

$$\mathcal{K}_+ = \left\{0, \ldots, \left\lfloor \frac{T}{2} \right\rfloor\right\}. \tag{13}$$

Let

$$K_+ = \left\lfloor \frac{T}{2} \right\rfloor + 1 \tag{14}$$

denote the number of nonnegative frequency indices. We define the boundaries

$$b_0 = 0, \qquad b_1 = \left\lceil \frac{K_+}{3} \right\rceil, \qquad b_2 = \left\lceil \frac{2K_+}{3} \right\rceil, \qquad b_3 = K_+. \tag{15}$$

and construct three contiguous frequency sets:

$$\begin{aligned} \mathcal{K}_{\text{low}} &= \{b_0, \ldots, b_1 - 1\}, \\ \mathcal{K}_{\text{mid}} &= \{b_1, \ldots, b_2 - 1\}, \\ \mathcal{K}_{\text{high}} &= \{b_2, \ldots, b_3 - 1\}. \end{aligned} \tag{16}$$

This construction assigns every nonnegative frequency coefficient to exactly one band while allowing the three bands to differ in size by at most one coefficient.

For each frequency band $f \in \mathcal{F}$, where

$$\mathcal{F} = \{\text{low}, \text{mid}, \text{high}\}, \tag{17}$$

we define a conjugate symmetric binary mask

$$M_f[k] = \begin{cases} 1, & k \in \mathcal{K}_f \text{ or } (-k \bmod T) \in \mathcal{K}_f, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The corresponding masked spectrum is

$$\mathbf{H}^f[k] = M_f[k]\mathbf{H}[k]. \tag{19}$$

The component associated with frequency band f is reconstructed using the inverse discrete Fourier transform (IDFT):

$$\tilde{\mathbf{x}}^{\mathrm{FM},f}[\tau] = \frac{1}{T}\sum_{k=0}^{T-1} \mathbf{H}^f[k] \exp\left(j\frac{2\pi k\tau}{T}\right). \tag{20}$$

The conjugate symmetric masks ensure that each reconstructed component is real valued. Since the masks form a disjoint partition of the complete spectrum,

$$\sum_{f\in\mathcal{F}} M_f[k] = 1 \tag{21}$$

for every frequency index k. Consequently, the original foundation model forecast can be recovered exactly:

$$\tilde{\mathbf{x}}^{\mathrm{FM}}[\tau] = \sum_{f\in\mathcal{F}} \tilde{\mathbf{x}}^{\mathrm{FM},f}[\tau]. \tag{22}$$

We apply the same decomposition to the corresponding ground-truth sequence when constructing the frequency-specific forecast errors and training objective. This produces band-specific targets $\mathbf{x}^f_{t:t+T-1}$ that are aligned with $\tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T-1}$.

A.2 Graph Residual Adapter Details

The Graph Residual Adapter performs frequency-specific residual correction without updating the parameters of the foundation model. For each $f \in \mathcal{F}$, the adapter receives the band-limited forecast $\tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T_{\mathrm{out}}-1}$ and the corresponding dynamic fusion prior graph $\mathbf{A}^{\mathrm{fusion}}_f$. Its processing flow consists of multi-scale patch embedding, graph propagation, temporal refinement, patch unembedding, and scale fusion.

For clarity, the frequency index f is omitted from intermediate token representations when no ambiguity arises.

A.2.1 Multi-Scale Time Patch Embedding

For each patch size $p \in \mathcal{P}$, we first transpose the band-limited forecast from $\mathbb{R}^{T_{\mathrm{out}}\times N}$ to $\mathbb{R}^{N\times T_{\mathrm{out}}}$, placing the station dimension first. We then divide its temporal dimension into nonoverlapping segments of length p. Let

$$T_p = \frac{T_{\mathrm{out}}}{p} \tag{23}$$

denote the resulting number of patch tokens. We define the patching operation as

$$\mathrm{Reshape}_p(\mathbf{X}) := \mathrm{reshape}_{N\times T_p\times p}\left(\mathbf{X}^\top\right). \tag{24}$$

Consequently,

$$\mathrm{Reshape}_p\left(\tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T_{\mathrm{out}}-1}\right) \in \mathbb{R}^{N\times T_p\times p}. \tag{25}$$

Each temporal patch is linearly projected into a D-dimensional token:

$$\mathbf{Z}^{(p)} = \mathrm{Reshape}_p\left(\tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T_{\mathrm{out}}-1}\right)\mathbf{W}^{\mathrm{in}}_p + \mathbf{b}^{\mathrm{in}}_p, \tag{26}$$

where $\mathbf{W}^{\mathrm{in}}_p \in \mathbb{R}^{p\times D}$ and $\mathbf{b}^{\mathrm{in}}_p \in \mathbb{R}^D$ are learnable parameters. The resulting representation satisfies

$$\mathbf{Z}^{(p)} \in \mathbb{R}^{N\times T_p\times D}. \tag{27}$$

We repeat this operation for every $p \in \mathcal{P}$. Different patch sizes provide complementary temporal resolutions: shorter patches preserve local fluctuations, whereas longer patches provide broader temporal context for residual correction.

A.2.2 Graph Normalization and Propagation

The dynamic fusion graph combines the static spatial graph and the frequency-specific error correlation graph:

$$\mathbf{A}^{\mathrm{fusion}}_f = \mathbf{A}^{\mathrm{spatial}} + \alpha_f \mathbf{A}^{\mathrm{error}}_f. \tag{28}$$

Although this graph supplies a frequency-specific station dependency prior, different stations may have different levels of receptivity to information from neighboring stations. We therefore refine the graph using a learnable node gating vector $\mathbf{g} \in \mathbb{R}^N$.

In implementation, the node gate is evaluated only on the nonzero edges of the sparsified fusion graph; the dense notation is used for clarity. The symmetric node gate is

$$\mathbf{G} = \mathrm{sig}\left(\mathbf{g}\mathbf{1}^\top + \mathbf{1}\mathbf{g}^\top\right), \tag{29}$$

where $\mathrm{sig}(\cdot)$ denotes the sigmoid function and $\mathbf{1} \in \mathbb{R}^N$ is an all-one vector. The gated adjacency matrix is

$$\overline{\mathbf{A}}^{\mathrm{fusion}}_f = \mathbf{A}^{\mathrm{fusion}}_f \odot \mathbf{G}. \tag{30}$$

Let

$$\mathbf{D}_f = \mathrm{diag}\left(\overline{\mathbf{A}}^{\mathrm{fusion}}_f \mathbf{1}\right) \tag{31}$$

be the corresponding row-degree matrix. Row normalization gives

$$\widetilde{\mathbf{A}}^{\mathrm{fusion}}_f = \mathbf{D}_f^{-1}\overline{\mathbf{A}}^{\mathrm{fusion}}_f. \tag{32}$$

This normalization controls the scale of the aggregated messages and ensures that each row sums to one.

Before propagation, each patch token is normalized over its D features for each sample, station, and patch. We denote its featurewise mean and standard deviation by $\boldsymbol{\mu}^{(p)}$ and $\boldsymbol{\sigma}^{(p)}$, respectively, and define $\mathrm{TokenNorm}(\mathbf{Z}^{(p)}) = (\mathbf{Z}^{(p)} - \boldsymbol{\mu}^{(p)})/\max(\boldsymbol{\sigma}^{(p)}, 10^{-5})$, where the maximum is applied elementwise. The statistics are retained to restore the token scale after temporal refinement. For every temporal token, graph propagation is performed along the station dimension:

$$\widetilde{\mathbf{Z}}^{(p)}_{:,\tau,:} = \widetilde{\mathbf{A}}^{\mathrm{fusion}}_f \,\mathrm{TokenNorm}(\mathbf{Z}^{(p)})_{:,\tau,:}, \qquad \tau = 1,\dots,T_p. \tag{33}$$

Therefore,

$$\widetilde{\mathbf{Z}}^{(p)} \in \mathbb{R}^{N\times T_p\times D}. \tag{34}$$

The propagation operation allows each station to aggregate residual information from stations connected by geographic relations, foundation model error dependencies, or both.

A.2.3 Time Attention and MLP

Graph propagation operates along the station dimension but does not explicitly model interactions among temporal patches. We therefore apply Time Attention independently to each station after graph propagation. At each patch scale, we first add a learnable positional embedding $\mathbf{P}^{(p)} \in \mathbb{R}^{1\times T_p\times D}$ to $\widetilde{\mathbf{Z}}^{(p)}$; the embeddings are initialized to zero. The temporal module applies pre-normalized attention and an MLP, each with a residual connection. Its output is rescaled using the token statistics retained before graph propagation.

For station i and attention head h, the query, key, and value representations are

$$\begin{aligned}\mathbf{Y}_i^{(p)} &= \mathrm{LayerNorm}\Big(\widetilde{\mathbf{Z}}_{i,:,:}^{(p)} + \mathbf{P}^{(p)}\Big),\\ \mathbf{Q}_i^{(p,h)} &= \mathbf{Y}_i^{(p)}\mathbf{W}_h^Q,\\ \mathbf{K}_i^{(p,h)} &= \mathbf{Y}_i^{(p)}\mathbf{W}_h^K,\\ \mathbf{V}_i^{(p,h)} &= \mathbf{Y}_i^{(p)}\mathbf{W}_h^V,\end{aligned} \tag{35}$$

where $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$, and $\mathbf{W}_h^V$ are learnable projection matrices.

The temporal attention output of head h is

$$\mathbf{U}_i^{(p,h)} = \mathrm{softmax}\left(\frac{\mathbf{Q}_i^{(p,h)}\left(\mathbf{K}_i^{(p,h)}\right)^\top}{\sqrt{d_h}}\right)\mathbf{V}_i^{(p,h)}, \tag{36}$$

where d_h is the dimension of each attention head. The outputs of all heads are concatenated and projected:

$$\mathbf{S}_i^{(p)} = \widetilde{\mathbf{Z}}_{i,:,:}^{(p)} + \mathbf{P}^{(p)} + \mathrm{Concat}\left(\mathbf{U}_i^{(p,1)},\ldots,\mathbf{U}_i^{(p,H)}\right)\mathbf{W}^O. \tag{37}$$

A lightweight MLP is subsequently applied to obtain

$$\begin{aligned}\mathbf{O}_i^{(p)} = \boldsymbol{\mu}_i^{(p)} &+ \max(\boldsymbol{\sigma}_i^{(p)}, 10^{-5})\odot\\ &\Big[\mathbf{S}_i^{(p)} + \mathrm{MLP}\Big(\mathrm{LayerNorm}(\mathbf{S}_i^{(p)})\Big)\Big].\end{aligned} \tag{38}$$

The refined tokens satisfy $\mathbf{O}^{(p)} \in \mathbb{R}^{N\times T_p\times D}$. Time Attention captures dependencies across temporal patches, whereas the MLP refines their feature representations before residual reconstruction.

For the 24-hour forecasting setting, the branch with $p = T_{\mathrm{out}} = 24$ contains a single token and therefore does not perform cross-token attention. Instead, this branch provides a global-window residual representation refined by the projection layers and MLP, while the shorter patch scales model interactions among multiple temporal tokens.

A.2.4 Patch Unembedding

The refined tokens are mapped back to temporal segments using a scale-specific output projection:

$$\mathbf{R}^{(p)} = \mathbf{O}^{(p)}\mathbf{W}_p^{\mathrm{out}} + \mathbf{b}_p^{\mathrm{out}}, \tag{39}$$

where $\mathbf{W}_p^{\mathrm{out}} \in \mathbb{R}^{D\times p}$ and $\mathbf{b}_p^{\mathrm{out}} \in \mathbb{R}^p$. This gives

$$\mathbf{R}^{(p)} \in \mathbb{R}^{N\times T_p\times p}. \tag{40}$$

The projected segments are rearranged into the original temporal resolution:

$$\Delta\tilde{\mathbf{x}}_{t:t+T_{\mathrm{out}}-1}^{(p,f)} = \mathrm{Reshape}_p^{-1}\left(\mathbf{R}^{(p)}\right), \tag{41}$$

where

$$\Delta\tilde{\mathbf{x}}_{t:t+T_{\mathrm{out}}-1}^{(p,f)} \in \mathbb{R}^{T_{\mathrm{out}}\times N}. \tag{42}$$

A.2.5 Dynamic Multi-Scale Fusion

The usefulness of a temporal scale may vary across samples and frequency bands. We therefore infer scale weights from the refined token representations rather than using fixed coefficients.

For each $p \in \mathcal{P}$, global average pooling over the station and token dimensions gives

$$\mathbf{c}^{(p)} = \frac{1}{NT_p}\sum_{i=1}^{N}\sum_{\tau=1}^{T_p}\mathbf{O}_{i,\tau,:}^{(p)} \in \mathbb{R}^D. \tag{43}$$

A lightweight scoring network produces a scalar score:

$$s^{(p)} = \mathrm{MLP}_{\mathrm{score}}\left(\mathbf{c}^{(p)}\right). \tag{44}$$

The scores are normalized across patch sizes:

$$\theta^{(p)} = \frac{\exp\left(s^{(p)}\right)}{\sum_{q\in\mathcal{P}} \exp\left(s^{(q)}\right)}. \tag{45}$$

The weights satisfy

$$\theta^{(p)} \geq 0, \qquad \sum_{p\in\mathcal{P}} \theta^{(p)} = 1. \tag{46}$$

The residual forecast for frequency band f is

$$\Delta\tilde{\mathbf{x}}^{f}_{t:t+T_{\mathrm{out}}-1} = \sum_{p\in\mathcal{P}} \theta^{(p)} \Delta\tilde{\mathbf{x}}^{(p,f)}_{t:t+T_{\mathrm{out}}-1}. \tag{47}$$

Finally, a learnable band factor β_f controls the correction magnitude of each frequency component:

$$\hat{\mathbf{x}}_{t:t+T_{\mathrm{out}}-1} = \sum_{f\in\mathcal{F}} \left(\tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T_{\mathrm{out}}-1} + \beta_f \Delta\tilde{\mathbf{x}}^{f}_{t:t+T_{\mathrm{out}}-1}\right). \tag{48}$$

This residual formulation preserves the temporal forecasts supplied by the frozen foundation model while introducing structured corrections based on station dependencies.

A.3 Training Objective

All foundation model parameters remain frozen during training. The trainable parameters include the Graph Residual Adapters, node gates $\mathbf{g}$, graph fusion factors α_f, band factors β_f, patch projections, Time Attention blocks, MLPs, and scale scorers.

A standard time-domain objective alone may be dominated by low-frequency components because they typically contain substantially more signal energy than high-frequency fluctuations. We therefore combine the MSE of the final forecast with an energy-normalized loss for each frequency band.

For frequency band f, the corrected prediction is

$$\hat{\mathbf{x}}^{f}_{t:t+T_{\mathrm{out}}-1} = \tilde{\mathbf{x}}^{\mathrm{FM},f}_{t:t+T_{\mathrm{out}}-1} + \beta_f \Delta\tilde{\mathbf{x}}^{f}_{t:t+T_{\mathrm{out}}-1}. \tag{49}$$

Let $\mathbf{x}^{f}_{t:t+T_{\mathrm{out}}-1}$ denote the corresponding ground-truth frequency component. The overall training objective is

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\mathrm{mse}}\,\mathrm{MSE}\left(\hat{\mathbf{x}}_{t:t+T_{\mathrm{out}}-1}, \mathbf{x}_{t:t+T_{\mathrm{out}}-1}\right) \\ & + \sum_{f\in\mathcal{F}} \lambda_f \sqrt{\frac{\left\|\hat{\mathbf{x}}^{f}_{t:t+T_{\mathrm{out}}-1} - \mathbf{x}^{f}_{t:t+T_{\mathrm{out}}-1}\right\|^2_{\mathrm{F}}}{\left\|\mathbf{x}^{f}_{t:t+T_{\mathrm{out}}-1}\right\|^2_{\mathrm{F}} + \varepsilon}},\end{aligned} \tag{50}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and $\varepsilon > 0$ is a small constant for numerical stability.

The denominator normalizes each band-specific error by the energy of its target component. This prevents high-energy low-frequency components from overwhelming the optimization and maintains meaningful gradients for the mid- and high-frequency adapters.

A.4 Complexity Analysis

We analyze the computational and memory complexity of WxFM-XL by separating offline graph construction from online adapter computation. Let N be the number of stations, T_{out} the forecasting horizon, D the token dimension, K the number of retained neighbors, and $T_p = T_{\mathrm{out}}/p$ the number of tokens associated with patch size p.

A.4.1 Comparison with Joint Spatiotemporal Attention

A direct approach can flatten the station and time dimensions into NT_{out} tokens and apply joint self-attention. Constructing the corresponding attention matrix requires

$$\mathcal{O}\left(N^2 T_{\text{out}}^2 D\right) \tag{51}$$

time and

$$\mathcal{O}\left(N^2 T_{\text{out}}^2\right) \tag{52}$$

memory.

A factorized architecture that applies dense spatial attention and temporal attention separately reduces the time complexity to

$$\mathcal{O}\left(N^2 T_{\text{out}} D + N T_{\text{out}}^2 D\right). \tag{53}$$

However, the spatial component remains quadratic in N, which becomes expensive for datasets containing thousands of stations.

A.4.2 Offline Prior Graph Construction

The spatial graph depends only on station coordinates and is computed once before training. Computing all pairwise station distances requires $\mathcal{O}(N^2)$ operations. The dense spatial similarities are then reduced to a sparse K-nearest-neighbor graph.

For each frequency band, the error correlation graph is constructed from the training errors. Let M denote the number of error instances after flattening the training sample and forecast lead-time dimensions. Computing the full Pearson correlation matrix requires

$$\mathcal{O}(MN^2) \tag{54}$$

operations. This computation is performed once for each foundation model and forecasting setting and does not contribute to the cost of online forecasting.

After sparsification, each graph contains at most $\mathcal{O}(KN)$ edges. The sparse graph storage cost over all frequency bands is

$$\mathcal{O}\left(|\mathcal{F}|KN\right). \tag{55}$$

A.4.3 Frequency Decomposition

For each station, computing the DFT over a forecast window of length T_{out} requires $\mathcal{O}(T_{\text{out}} \log T_{\text{out}})$ operations. The DFT over all stations therefore costs $\mathcal{O}(NT_{\text{out}} \log T_{\text{out}})$. Reconstructing the $|\mathcal{F}|$ frequency components using the IDFT requires $\mathcal{O}(|\mathcal{F}|NT_{\text{out}} \log T_{\text{out}})$ operations. Thus, the total cost of frequency decomposition is

$$\mathcal{O}((1 + |\mathcal{F}|)NT_{\text{out}} \log T_{\text{out}}). \tag{56}$$

Storing all band-limited components requires $\mathcal{O}(|\mathcal{F}|NT_{\text{out}})$ memory.

A.4.4 Multi-Scale Patch Embedding

For patch size p, patch embedding transforms NT_p temporal patches, each containing p values, into D-dimensional tokens. Its complexity is

$$\mathcal{O}\left(NT_p p D\right) = \mathcal{O}\left(NT_{\text{out}} D\right). \tag{57}$$

Patch unembedding has the same asymptotic cost.

A.4.5 Graph Normalization and Propagation

The node gate and row normalization operate only on the retained sparse edges and require $\mathcal{O}(KN)$ operations. Propagating T_p tokens with dimension D over the sparse graph requires

$$\mathcal{O}\left(KNT_p D\right) \tag{58}$$

operations.

This avoids dense spatial message passing, whose complexity would be $\mathcal{O}(N^2 T_p D)$.

A.4.6 Temporal Refinement

Time Attention is applied independently to the T_p tokens of each station. Constructing the temporal attention maps requires

$$\mathcal{O}\left(NT_p^2 D\right) \tag{59}$$

operations. The query, key, value, output projections, and MLP require

$$\mathcal{O}\left(NT_p D^2\right) \tag{60}$$

operations.

Compared with applying temporal attention directly to the T_{out} original time steps, patching reduces the quadratic attention term by a factor of p^2 because $T_p = T_{\text{out}}/p$.

A.4.7 Scale Scoring and Fusion

Global pooling requires $\mathcal{O}(NT_p D)$ operations for each patch scale. The scale scorer operates on a D-dimensional context vector, and its cost is independent of N. Reconstructing and combining the scale-specific residual forecasts requires $\mathcal{O}(NT_{\text{out}})$ operations.

A.4.8 Overall Online Complexity

Combining the above components, the online complexity of the adapter is

$$\mathcal{O}\left((1+|\mathcal{F}|)NT_{\text{out}} \log T_{\text{out}} + |\mathcal{F}| \sum_{p\in\mathcal{P}} NT_p \left(KD + T_p D + D^2 + pD\right)\right). \tag{61}$$

The corresponding activation memory is dominated by the patch tokens and temporal attention maps:

$$\mathcal{O}\left(|\mathcal{F}| \sum_{p\in\mathcal{P}} N\left(T_p D + T_p^2\right)\right). \tag{62}$$

The sparse prior graphs additionally require $\mathcal{O}(|\mathcal{F}|KN)$ memory. The band-limited forecasts additionally require $\mathcal{O}(|\mathcal{F}|NT_{\text{out}})$ memory.

Since $|\mathcal{F}|$, $|\mathcal{P}|$, K, D, and T_p are fixed with respect to the number of stations, both the online computation and activation memory grow linearly with N. The proposed adapter therefore avoids the quadratic station dependence of joint or dense spatial attention and remains applicable to large multi-station forecasting datasets.

B Experimental Details

This section provides the information required to reproduce the experiments. We first describe the three datasets and their chronological splits, then summarize the baseline configurations, training settings, random-seed protocol, and evaluation metrics.

B.1 Dataset Details

We conduct experiments on three station-based weather datasets with different geographical scales and station distributions.

B.1.1 Hunan Regional Dataset

The Hunan regional dataset is provided by the Hunan Province Meteorological Bureau. It contains hourly observations from 3,570 weather stations distributed across an approximately $576\,\text{km} \times 576\,\text{km}$ region in Hunan Province, China.

The dataset covers the period from April 1 to September 30, 2023. Each station provides temperature at 2 meters and the u- and v-components of wind at 10 meters. The large number of densely distributed stations makes this dataset suitable for evaluating scalability and fine-grained regional station dependencies.

B.1.2 FRENCH REGIONAL DATASET

The French regional dataset is obtained from MeteoNet (Larvor et al., 2020), which is released by Meteo France. It contains hourly observations from 234 weather stations distributed within an approximately $252\,\mathrm{km} \times 252\,\mathrm{km}$ region in northwestern France.

The dataset spans from January 1 to December 31, 2016. The evaluated variables are temperature at 2 meters and the u- and v-components of wind at 10 meters. Compared with the Hunan dataset, the French dataset contains fewer and more sparsely distributed stations and therefore provides a complementary regional forecasting setting.

B.1.3 GLOBAL DATASET

The global dataset is obtained from the National Centers for Environmental Information (NOAA National Centers for Environmental Information, 2026). It contains hourly observations from 3,850 stations distributed worldwide between January 1, 2019, and December 31, 2020.

The evaluated variables are hourly temperature and averaged wind speed. The global distribution of stations introduces substantially larger spatial variation than the two regional datasets. Because the Global variables were scaled by a factor of 10 during preprocessing, we divide the resulting MSE by 100 and MAE by 10 before reporting them.

B.1.4 DATASET SPLITS

For every dataset and variable, observations are arranged chronologically and divided into training, validation, and test sets using a 6:1:3 ratio. Random shuffling across time is not used.

The training split is used to optimize trainable models and construct the frequency-specific error correlation prior graphs. The validation split is used for model selection, and all reported results are computed on the held-out test split. Neither the validation nor test targets are used when constructing the error correlation prior graphs.

B.2 BASELINE DETAILS

We compare WxFM-XL with three groups of forecasting methods. The grouping follows the organization used in the Related Work and experimental sections.

B.2.1 TIME SERIES FOUNDATION MODELS

Timer (Liu et al., 2024) is a generative pretrained time series model that formulates forecasting through next-token prediction. Sundial (Liu et al., 2025b) performs probabilistic forecasting over continuous-valued observations using flow matching.

We evaluate their officially released pretrained models without updating their parameters on the target weather datasets. These experiments measure the zero-shot performance of univariate foundation models when each station is forecast independently.

B.2.2 FOUNDATION MODELS FOR MULTIVARIATE FORECASTING

Timer-XL (Liu et al., 2025a) extends Timer to multivariate time series and models dependencies among variables through TimeAttention. Moirai (Woo et al., 2024) is a universal forecasting model designed to support heterogeneous frequencies and different numbers of variables.

For Timer-XL and Moirai, we report both zero-shot and fine-tuned results. Their fine-tuned variants are denoted as Timer-XL-ft and Moirai-ft, respectively.

AdaPTS (Benechehab et al., 2025) adapts pretrained univariate foundation models to multivariate inputs through lightweight projection adapters. We combine AdaPTS with Timer and Sundial and denote the resulting variants as AdaPTS (Timer) and AdaPTS (Sundial). The adapters are trained separately on each target dataset.

B.2.3 MULTI-STATION FORECASTING MODELS

CDPNet (Xu et al., 2025) models continuous spatiotemporal evolution from discrete station observations. MSGNet (Cai et al., 2024) combines multiscale frequency decomposition with adaptive graph convolution. Corrformer (Wu et al., 2023) integrates spatial cross-correlation and temporal autocorrelation for weather forecasting.

STELLA (Fu et al., 2025) incorporates geographic coordinates and temporal features through spatial-temporal position embedding and uses a lightweight MLP architecture instead of Transformer layers. These task-specific models are trained on each target dataset using their official or recommended configurations.

B.3 IMPLEMENTATION DETAILS

Whenever available, all baselines follow their official implementations and recommended hyperparameter settings. Timer, Sundial, Timer-XL, and Moirai use their officially released pretrained weights for zero-shot evaluation. Timer-XL-ft and Moirai-ft are trained on each target dataset using their official fine-tuning scripts.

For AdaPTS, we evaluate its available adapter configurations and use the dropout Linear Adapter Ensemble configuration with both Timer and Sundial. This configuration provides the strongest overall validation performance across our experimental settings.

We instantiate WxFM-XL using frozen pretrained Timer and Sundial backbones. These two variants are denoted as WxFM-XL (Timer) and WxFM-XL (Sundial), respectively. During training, the foundation model parameters remain frozen, and only the adapter, patch projections, graph gates, fusion factors, temporal modules, and scale scorers are updated.

The input window is fixed at 96 hours to accommodate the input length required by Timer. The same input window is used for every method to ensure a consistent comparison. We evaluate forecast horizons

$$T_{\text{out}} \in \{24, 96\}. \tag{63}$$

The number of frequency bands is fixed to three:

$$\mathcal{F} = \{\text{low}, \text{mid}, \text{high}\}. \tag{64}$$

The temporal patch sizes are

$$\mathcal{P} = \{6, 12, 24\}. \tag{65}$$

The token dimension is $D = 128$. Time Attention uses four heads of dimension 32 and an attention dropout rate of 0.1. The following MLP has a hidden dimension of 256 and a dropout rate of 0.1. The scale-weight scorer has a hidden dimension of 32.

Both the spatial correlation graph and each frequency-specific error correlation prior graph use the same dataset-specific neighborhood size: $k = 32$ for Hunan, $k = 16$ for MeteoNet (French regional), and $k = 36$ for Global. Self-loops are included in both graph types. The training objective uses $\lambda_{\text{mse}} = 1.0$ and $\lambda_f = 0.5$ for every $f \in \mathcal{F}$.

We optimize WxFM-XL using Adam with an initial learning rate of 1×10^{-4}, which is annealed to 1×10^{-6} using a cosine schedule, for up to 20 epochs. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory.

Stochastic experiments are repeated using random seeds 2021, 2022, and 2023. For each metric, we report the arithmetic mean and sample standard deviation across the three runs.

The zero-shot evaluations of Timer and Timer-XL use deterministic inference and therefore exhibit no variation across random seeds. Although Sundial and Moirai are not strictly deterministic, their observed standard deviations under our evaluation protocol are close to zero and negligible at the reported precision. We therefore report these four foundation models without standard deviations. For all trainable models, results are reported as the mean and sample standard deviation over three random seeds.

B.4 Evaluation Metrics

We evaluate forecasting accuracy using mean squared error (MSE) and mean absolute error (MAE). Lower values indicate better performance for both metrics.

Let Ω denote the set of evaluated sample, forecast-step, and station indices. MSE is defined as

$$\mathrm{MSE} = \frac{1}{|\Omega|} \sum_{(s,\tau,i)\in\Omega} \left(\hat{x}_{s,\tau,i} - x_{s,\tau,i}\right)^2, \tag{66}$$

and MAE is defined as

$$\mathrm{MAE} = \frac{1}{|\Omega|} \sum_{(s,\tau,i)\in\Omega} \left|\hat{x}_{s,\tau,i} - x_{s,\tau,i}\right|. \tag{67}$$

Let m_r denote the metric obtained using the r-th random seed. For $R = 3$ runs, the reported mean is

$$\bar{m} = \frac{1}{R} \sum_{r=1}^{R} m_r, \tag{68}$$

and the sample standard deviation is

$$s_m = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} \left(m_r - \bar{m}\right)^2}. \tag{69}$$

The standard deviation therefore uses Bessel's correction, corresponding to `ddof=1` in numerical implementations.

B.4.1 Weighted Edge Overlap

For Fig. 4(c), we compare the spatial and error prior graphs after top-k sparsification and symmetrization. We exclude self-loops and assign zero weight to absent edges. For two graphs A and B, let $\mathcal{E}_{AB} = \{(i,j) : i < j,\ A_{ij} \text{ and } B_{ij} \text{ are finite}\}$ and $w_{ij}^A = \max(A_{ij}, 0)$, $w_{ij}^B = \max(B_{ij}, 0)$. Their weighted edge overlap (weighted Jaccard index) is

$$J(A,B) = \frac{\sum_{(i,j)\in\mathcal{E}_{AB}} \min(w_{ij}^A, w_{ij}^B)}{\sum_{(i,j)\in\mathcal{E}_{AB}} \max(w_{ij}^A, w_{ij}^B)}. \tag{70}$$

The overlap is undefined if the denominator is zero.

C Additional Experimental Results

This section supplements the main experiments with complete MAE results and qualitative forecasting cases. The MAE results cover all datasets, variables, and forecasting horizons under the same experimental settings as the MSE results reported in the main paper. The forecasting cases further illustrate the behavior of different methods on representative weather sequences.

C.1 Complete MAE Results

Table 3 reports the complete MAE results on the Hunan regional, French regional, and global datasets for forecasting horizons of 24 and 96 hours. For trainable models, we report the mean and sample standard deviation over three runs. Timer and Timer-XL use deterministic inference. Sundial and Moirai are not strictly deterministic, but their observed variations are negligible relative to their mean values; their standard deviations are therefore omitted. Best and second-best results are marked in bold and underlined, respectively, based on the unrounded means.

Table 4 provides the complete MAE results corresponding to the MSE ablation study in the main paper. The complete model achieves the lowest MAE across all variables on both regional datasets, confirming that the improvements observed in MSE are also consistent under the absolute-error metric. Removing either the frequency domain decomposition or the error correlation prior graph

Table 3: Complete MAE results on the Hunan regional, French regional, and global datasets with a 96-hour input window and forecast horizons of 24 and 96 hours. Results are reported as the mean and sample standard deviation over three random seeds where applicable. Best and second-best results are shown in **bold** and underlined, respectively.

Models			WxFM-XL (Timer)	WxFM-XL (Sundial)	Timer	Sundial	AdaPTS (Timer)	AdaPTS (Sundial)	$\text{Moirai}_{\text{ft}}$	Moirai	$\text{Timer-XL}_{\text{ft}}$	Timer-XL	STELLA	CDPNet	MSGNet	Corrformer
Hunan	temp	24	**1.552**$_{\pm 0.031}$	1.586$_{\pm 0.005}$	1.627	1.679	1.661$_{\pm 0.055}$	1.735$_{\pm 0.056}$	1.715$_{\pm 0.103}$	1.590	1.697$_{\pm 0.015}$	1.662	2.236$_{\pm 0.131}$	2.023$_{\pm 0.111}$	1.760$_{\pm 0.110}$	1.841$_{\pm 0.100}$
		96	1.952$_{\pm 0.016}$	**1.947**$_{\pm 0.005}$	1.987	2.119	2.094$_{\pm 0.078}$	2.861$_{\pm 0.127}$	1.981$_{\pm 0.019}$	1.991	2.269$_{\pm 0.041}$	1.966	3.100$_{\pm 0.181}$	4.027$_{\pm 0.564}$	2.680$_{\pm 0.095}$	2.256$_{\pm 0.068}$
	v-wind	24	**0.737**$_{\pm 0.003}$	0.761$_{\pm 0.002}$	0.745	0.817	0.751$_{\pm 0.032}$	0.763$_{\pm 0.020}$	0.766$_{\pm 0.012}$	0.771	0.794$_{\pm 0.002}$	0.766	0.759$_{\pm 0.011}$	0.797$_{\pm 0.013}$	0.775$_{\pm 0.054}$	0.882$_{\pm 0.064}$
		96	**0.842**$_{\pm 0.004}$	0.931$_{\pm 0.028}$	0.850	0.899	0.875$_{\pm 0.040}$	0.882$_{\pm 0.044}$	0.891$_{\pm 0.039}$	0.898	0.847$_{\pm 0.040}$	0.846	0.888$_{\pm 0.045}$	1.114$_{\pm 0.335}$	1.001$_{\pm 0.039}$	0.949$_{\pm 0.073}$
	u-wind	24	0.689$_{\pm 0.002}$	**0.661**$_{\pm 0.018}$	0.676	0.751	0.698$_{\pm 0.023}$	0.712$_{\pm 0.016}$	0.711$_{\pm 0.003}$	0.710	0.698$_{\pm 0.003}$	0.686	0.674$_{\pm 0.010}$	0.711$_{\pm 0.017}$	0.719$_{\pm 0.038}$	0.676$_{\pm 0.045}$
		96	0.696$_{\pm 0.004}$	**0.693**$_{\pm 0.018}$	0.709	0.795	0.717$_{\pm 0.022}$	0.740$_{\pm 0.011}$	0.767$_{\pm 0.005}$	0.778	0.699$_{\pm 0.003}$	0.720	0.707$_{\pm 0.006}$	0.769$_{\pm 0.038}$	0.792$_{\pm 0.042}$	0.724$_{\pm 0.050}$
French	temp	24	1.943$_{\pm 0.011}$	1.927$_{\pm 0.011}$	2.071	2.107	2.013$_{\pm 0.087}$	2.073$_{\pm 0.042}$	1.971$_{\pm 0.011}$	2.032	**1.925**$_{\pm 0.037}$	2.009	2.230$_{\pm 0.031}$	3.703$_{\pm 0.307}$	2.013$_{\pm 0.096}$	2.452$_{\pm 0.078}$
		96	**2.692**$_{\pm 0.026}$	2.771$_{\pm 0.022}$	2.721	2.809	2.754$_{\pm 0.035}$	2.770$_{\pm 0.052}$	2.743$_{\pm 0.127}$	2.734	2.850$_{\pm 0.201}$	2.768	3.380$_{\pm 0.161}$	7.446$_{\pm 0.104}$	2.843$_{\pm 0.222}$	2.824$_{\pm 0.073}$
	v-wind	24	**1.487**$_{\pm 0.006}$	1.498$_{\pm 0.002}$	1.655	1.724	1.625$_{\pm 0.065}$	1.582$_{\pm 0.075}$	1.592$_{\pm 0.016}$	1.696	1.663$_{\pm 0.007}$	1.663	1.654$_{\pm 0.014}$	1.596$_{\pm 0.051}$	1.573$_{\pm 0.114}$	1.731$_{\pm 0.134}$
		96	**1.821**$_{\pm 0.005}$	1.868$_{\pm 0.072}$	2.007	2.147	1.943$_{\pm 0.115}$	1.863$_{\pm 0.044}$	2.043$_{\pm 0.014}$	2.279	2.051$_{\pm 0.036}$	2.009	1.981$_{\pm 0.057}$	1.962$_{\pm 0.077}$	2.140$_{\pm 0.163}$	2.071$_{\pm 0.096}$
	u-wind	24	**1.556**$_{\pm 0.007}$	1.560$_{\pm 0.001}$	1.820	1.934	1.624$_{\pm 0.024}$	1.700$_{\pm 0.024}$	1.686$_{\pm 0.011}$	1.758	1.583$_{\pm 0.008}$	1.854	1.776$_{\pm 0.018}$	1.648$_{\pm 0.139}$	1.617$_{\pm 0.066}$	1.809$_{\pm 0.109}$
		96	**2.148**$_{\pm 0.006}$	2.149$_{\pm 0.028}$	2.327	2.524	2.169$_{\pm 0.114}$	2.154$_{\pm 0.026}$	2.350$_{\pm 0.032}$	2.562	2.188$_{\pm 0.050}$	2.336	2.242$_{\pm 0.038}$	2.254$_{\pm 0.051}$	2.320$_{\pm 0.130}$	2.303$_{\pm 0.114}$
Global	temp	24	1.903$_{\pm 0.003}$	1.881$_{\pm 0.004}$	2.143	2.186	2.050$_{\pm 0.086}$	2.034$_{\pm 0.120}$	2.052$_{\pm 0.003}$	2.109	1.949$_{\pm 0.036}$	2.337	**1.861**$_{\pm 0.003}$	2.126$_{\pm 0.020}$	2.251$_{\pm 0.116}$	1.946$_{\pm 0.063}$
		96	**2.625**$_{\pm 0.010}$	2.635$_{\pm 0.040}$	2.886	3.003	2.788$_{\pm 0.088}$	2.857$_{\pm 0.093}$	2.785$_{\pm 0.030}$	2.998	2.757$_{\pm 0.015}$	3.089	2.723$_{\pm 0.014}$	2.725$_{\pm 0.065}$	3.024$_{\pm 0.031}$	2.914$_{\pm 0.048}$
	wind	24	1.282$_{\pm 0.001}$	**1.278**$_{\pm 0.001}$	1.377	1.494	1.375$_{\pm 0.051}$	1.398$_{\pm 0.011}$	1.371$_{\pm 0.004}$	1.449	1.368$_{\pm 0.013}$	1.405	1.290$_{\pm 0.001}$	1.408$_{\pm 0.011}$	1.564$_{\pm 0.057}$	1.366$_{\pm 0.094}$
		96	1.488$_{\pm 0.036}$	1.500$_{\pm 0.024}$	1.563	1.691	1.532$_{\pm 0.072}$	1.544$_{\pm 0.055}$	1.560$_{\pm 0.007}$	1.763	1.497$_{\pm 0.004}$	1.578	**1.476**$_{\pm 0.004}$	1.532$_{\pm 0.027}$	1.802$_{\pm 0.008}$	1.559$_{\pm 0.051}$

Table 4: Complete MAE results of the ablation study on the Hunan regional and French regional datasets. The input window is 96 hours, and the forecast horizon is 24 hours. Results are reported as the mean and sample standard deviation over three random seeds. Best results are shown in **bold**.

Model variant	Hunan regional			French regional		
	temp	v-wind	u-wind	temp	v-wind	u-wind
WxFM-XL (Timer)	**1.552**$_{\pm 0.031}$	**0.737**$_{\pm 0.003}$	**0.689**$_{\pm 0.002}$	**1.943**$_{\pm 0.011}$	**1.487**$_{\pm 0.006}$	**1.556**$_{\pm 0.007}$
w/o frequency domain decomposition	1.682$_{\pm 0.041}$	0.791$_{\pm 0.008}$	0.712$_{\pm 0.016}$	1.985$_{\pm 0.020}$	1.586$_{\pm 0.016}$	1.641$_{\pm 0.004}$
w/o error correlation prior graph	1.684$_{\pm 0.038}$	0.793$_{\pm 0.003}$	0.713$_{\pm 0.004}$	1.977$_{\pm 0.022}$	1.619$_{\pm 0.004}$	1.636$_{\pm 0.018}$
w/o multi-scale temporal module	1.753$_{\pm 0.062}$	0.797$_{\pm 0.008}$	0.714$_{\pm 0.010}$	2.002$_{\pm 0.014}$	1.652$_{\pm 0.010}$	1.737$_{\pm 0.015}$

consistently degrades the forecasting accuracy. The variant without the multi-scale temporal module exhibits the largest overall degradation, particularly for temperature and wind forecasting on the French regional dataset. These results further demonstrate that frequency-specific correction, error-aware spatial propagation, and multi-scale temporal modeling contribute complementary improvements.

C.2 Forecasting Case Studies

To complement the aggregate evaluation metrics, we present representative forecasting cases selected from the test sets. The cases cover different datasets, weather variables, and forecasting horizons, including both smooth temporal trends and rapid local fluctuations. For each case, we compare the ground-truth observations with the predictions produced by the frozen foundation model and WxFM-XL.

These comparisons provide a direct view of how the proposed adapter modifies the original foundation model forecasts. In particular, they allow us to examine whether WxFM-XL can correct systematic forecast deviations, preserve the overall temporal trend, and capture local variations across correlated stations. We also include long-horizon cases to investigate the accumulation of forecasting errors as the prediction horizon increases.

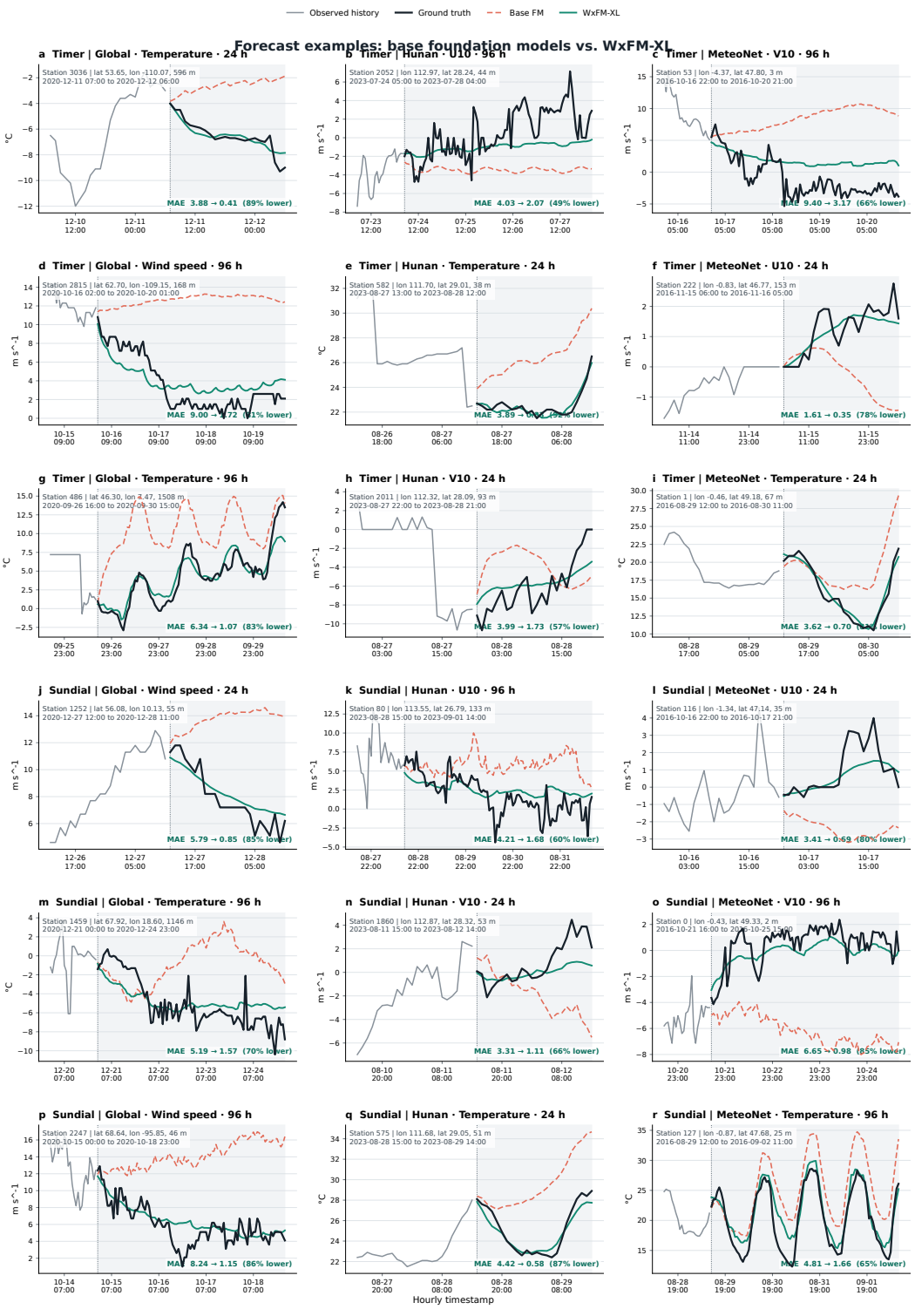


Figure 5: Representative forecasting cases on the evaluated datasets. The plots compare the ground truth with predictions from the frozen foundation model and WxFM-XL.