

Article

TF-STNet: A Time–Frequency Dual-Branch Spatiotemporal Network for NWP-to-Station Bias Correction

Zhao Wang ¹, Shuai Chen ¹, Yunbo Yang ¹, Bo Wang ¹, Bihe Xu ² and Yangliao Geng ^{2,*}

¹ State Key Laboratory of Renewable Energy Grid-Integration, China Electric Power Research Institute, Beijing 100192, China; wangzhao@epri.sgcc.com.cn (Z.W.); chenshuai@epri.sgcc.com.cn (S.C.); yangyunbo@epri.sgcc.com.cn (Y.Y.); wangbo@epri.sgcc.com.cn (B.W.)

² Key Laboratory of Big Data & Artificial Intelligence in Transportation (Ministry of Education), Beijing Jiaotong University, Beijing 100044, China; xbhnew@bjtu.edu.cn

* Correspondence: ylageng@bjtu.edu.cn

Abstract

Accurate station-scale weather forecasts support renewable-energy and transportation operations, yet gridded numerical weather prediction (NWP) is affected by representativeness errors and regional biases when transferred to irregularly distributed stations. We propose TF-STNet, a time–frequency dual-branch spatiotemporal network for NWP-to-station bias correction. A station-centered K-nearest-neighbor (KNN) operator retains multiple local NWP trajectories. The time-domain pathway separately encodes recent observations and future NWP, aligns them over the forecast horizon using gated dilated causal convolutions, and propagates lead-resolved states through a coordinate-conditioned directed station graph. The frequency-domain pathway learns spectral weights and applies separate attention to amplitude and phase across neighboring NWP cells. Prediction-level fusion combines the two station forecasts by variable, station, and lead time. The evaluation uses hourly data for wind speed, pressure, relative humidity, and temperature from 455 stations in Hebei, Shandong, Fujian, and Sichuan. Across five independent runs on 16 region–variable tasks, TF-STNet achieves the lowest mean absolute error (MAE) in 15 tasks and the highest Pearson correlation coefficient (PCC) in 15 tasks; its pressure MAE reduction relative to the strongest learned comparator ranges from 16.2% to 57.4% across the four regions. It has lower MAE than raw NWP in seven of eight high-wind or rapid-change event tests and than simple pressure model-output-statistics corrections in all four regions. The Shandong–temperature task and the Hebei–high-wind case illustrate the limits of the present point-forecast formulation.



Academic Editors: Jaroslaw Krzywanski, Marcin Sosnowski, Karolina Grabowska, Dorian Skrobek, Anna Zylka, Agnieszka Kijo-Kleczkowska, Bashar Shboul and Tomasz Czakiert

Received: 13 August 2026

Revised: 4 September 2026

Accepted: 4 September 2026

Published: 8 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: numerical weather prediction; spatiotemporal learning; time–frequency representation; graph neural networks; station-scale forecasting; meteorological post-processing; spectral attention

1. Introduction

Accurate station-scale forecasting is a fundamental task in meteorology, supporting weather monitoring, early warning, environmental management, and a wide range of weather-sensitive activities; recent limited-area AI models further underscore the demand for high-resolution forecasts of surface meteorological variables [1]. Modern observation networks and numerical weather prediction (NWP) systems generate large, heterogeneous spatiotemporal datasets, creating substantial opportunities for artificial intelligence and neural networks to learn nonlinear relationships across variables, locations, and forecast lead

times, as demonstrated by large-scale Earth-system foundation models [2]. NWP remains indispensable because it provides physically consistent guidance throughout the forecast horizon. However, its grid cells represent area-averaged atmospheric states, whereas station observations describe specific locations. Differences in elevation, land cover, coastal exposure, and unresolved local processes consequently produce systematic and time-varying discrepancies between gridded forecasts and station measurements. Conventional spatial interpolation can reduce this mismatch in spatial support, but it cannot learn recurring grid-to-station relationships or jointly represent their temporal evolution and spatial dependence; recent neural downscaling studies demonstrate the value of learned mappings for reconstructing fine-scale meteorological structure from coarse fields [3].

Statistical post-processing provides a direct way to estimate such forecast–observation relationships from historical data. Classical model output statistics and ensemble post-processing established this principle, while neural post-processing extended it to nonlinear and multivariate mappings for both deterministic and probabilistic products [4,5]. Recent work has also formulated coarse-to-fine rainfall downscaling as conditional distribution learning with spatial dependence [6]. Nevertheless, performance remains sensitive to the forecasting method and to changes in nonlinearity, dependence, and temporal complexity [7]. These challenges are especially pronounced in regional meteorology: wind speed, pressure, relative humidity, and temperature differ in scale, persistence, periodicity, and terrain sensitivity, and their NWP errors vary across coastal, plain, plateau, and basin environments. An effective correction model must therefore accommodate multiple forms of heterogeneity without allowing one variable or information source to dominate the learned representation.

Meanwhile, data-driven global weather forecasting has advanced rapidly. FourCastNet introduced adaptive Fourier neural operators for high-resolution global prediction [8]; Pangu-Weather, GraphCast, and FuXi demonstrated skillful medium-range forecasting through three-dimensional neural networks, graph-based message passing, and cascaded lead-time models [9–11]; and generative systems subsequently extended these advances to probabilistic forecasting [12]. Hybrid neural general circulation models combine learned components with differentiable physical solvers [13], while WeatherBench 2 provides a standardized framework for comparing global systems [14]. These models offer broad skill and rapid inference, but they reconstruct atmospheric states on regular grids and are optimized for large-scale performance rather than the terrain, exposure, and measurement characteristics of individual stations. ASTAFN directly addresses this gap by using proxy-station learning to align weather foundation model outputs with station observations and adaptively fuse their spatiotemporal features [15]. More generally, model rankings vary with the target variable, lead time, and verification regime [16]. Thus, advances in global forecasting strengthen the physically grounded guidance available for local prediction but do not eliminate the need for dedicated station-level correction.

At the station scale, NWP correction has a distinctive spatiotemporal structure. Historical observations describe the current station state and its recent local discrepancy, whereas NWP is the only input that spans the future forecast horizon. The two sources are therefore temporally adjacent but statistically asymmetric: one is measured and retrospective, while the other is model-generated and prospective. In addition, weather systems couple irregularly distributed stations, and these dependencies may be directional or shaped by factors beyond geographic distance. Graph WaveNet learns latent adjacency jointly with temporal dynamics [17], MegaCRN represents recurring spatial patterns through meta-graphs [18], and MTGNN infers multivariate relations from data [19]. Gated temporal representations can also benefit from learned graph structure when spatial dependencies evolve [20]. However, these graph forecasters are generally formulated for node-aligned temporal inputs and do not explicitly reconcile measured history with gridded future guidance.

NWP correction residuals also exhibit complementary structure in the time and frequency domains. Temporal order and short-range context are essential to representing fronts, ramps, and abrupt local transitions, whereas recurrent or slowly varying errors across the full forecast horizon are often more directly expressed through spectral content. Attention-based forecasters capture long-range dependence [21,22]; PatchTST learns patch-wise, channel-independent representations [23]; DLinear demonstrates the strength of decomposed linear structure [24]; and Temporal Fusion Transformers incorporate heterogeneous covariates in multi-horizon forecasting [25]. Frequency-aware models such as Autoformer, FEDformer, TimesNet, and FreTS further demonstrate the value of periodic and multi-scale representations [26–29], while PHAT explicitly models heterogeneous and dynamically changing periods across variables [30]. More broadly, time–frequency analysis can reveal signal organization that is difficult to isolate in either domain alone [31]. This distinction is particularly relevant to bias correction: an NWP trajectory may reproduce a diurnal cycle but misestimate its amplitude, or it may capture the strength of a change while shifting its arrival time. Although both cases produce pointwise errors, they require different forms of correction.

Taken together, NWP-to-station correction requires resolving three coupled mismatches: measured history must be aligned with model-generated future guidance, regular NWP grids must be translated to irregular station locations, and lead-specific temporal changes must be distinguished from horizon-scale spectral errors. Existing graph and frequency-aware forecasters address individual aspects of this problem but generally assume that their inputs already share common temporal and spatial support. They also do not explicitly distinguish whether neighboring NWP trajectories are reliable in oscillation amplitude, timing, or both. Naively fusing all inputs at an early stage would therefore force heterogeneous temporal axes, spatial structures, and feature semantics into one latent representation, potentially weakening source-specific learning and obscuring abrupt lead-dependent changes. These considerations motivate an architecture that preserves source- and domain-specific representations until they can be combined at a common station-forecast level.

To address these limitations, this paper introduces TF-STNet, a time–frequency dual-branch spatiotemporal network that preserves source and representation boundaries until meaningful alignment is possible. A shared station-centered KNN operator retains multiple local NWP trajectories instead of collapsing them into a single interpolated value. The time-domain pathway separately encodes observation history and future NWP, projects the historical state onto the forecast horizon, and propagates the fused lead-resolved features through a coordinate-conditioned directed station graph. In parallel, the frequency-domain pathway learns the relevance of spectral components and applies separate amplitude and phase attention across neighboring NWP cells, allowing the preferred spatial evidence to vary by station and oscillatory scale. Each pathway first produces a station-level forecast; prediction-level fusion then combines outputs with identical physical dimensions and adjusts their contributions by variable, station, and lead time. The resulting pathways act as complementary correction estimators: the time-domain pathway emphasizes causal local evolution and inter-station propagation, whereas the frequency-domain pathway targets horizon-scale distortions in oscillatory strength and timing.

The main contributions are as follows:

1. We propose TF-STNet, a time–frequency fusion framework that jointly captures lead-specific local departures and horizon-scale distortions in the amplitude and timing of NWP variations. By combining complementary station-level estimates only after each representation has specialized, the framework avoids premature feature mixing and adapts its correction to each variable, station, and forecast lead.

- TF-STNet introduces a station-centered amplitude–phase attention mechanism that processes Fourier amplitude and phase separately across neighboring NWP cells. This mechanism distinguishes errors in oscillatory strength from timing shifts and selects the most informative local evidence for each component.
- Experiments across 455 stations, four meteorological variables, and four geographically distinct Chinese regions yield 16 region–variable tasks. Across five independent runs, TF-STNet achieves the lowest mean MAE and the highest mean PCC in 15 of the 16 tasks; Shandong–temperature is the only task for which another method has a lower mean MAE.

2. Problem Formulation and Proposed Method

2.1. Problem Definition

Let a forecast be issued at time t for N target stations and C meteorological variables. The available inputs are a recent sequence of station observations and a gridded NWP trajectory over the future forecast horizon; the target is the corresponding future station sequence. Throughout this section, bold symbols denote vectors, matrices, or higher-order tensors; subscripts index tensor axes, while descriptive superscripts label representations rather than powers. Omitting the batch dimension, the principal tensors are

$$\mathbf{X} \in \mathbb{R}^{T_{\text{in}} \times N \times C}, \quad \mathbf{E} \in \mathbb{R}^{T_{\text{out}} \times N_{\text{lat}} \times N_{\text{lon}} \times C}, \quad \mathbf{Y} \in \mathbb{R}^{T_{\text{out}} \times N \times C}, \quad (1)$$

where T_{in} and T_{out} are the numbers of hourly input and output steps, respectively; both equal 24 in this study. The number of variables is $C = 4$, corresponding to wind speed, pressure, relative humidity, and temperature, while N_{lat} and N_{lon} denote the numbers of NWP grid points along the latitude and longitude axes. Station and NWP-grid coordinates are denoted by $\mathbf{S} \in \mathbb{R}^{N \times 2}$ and $\mathbf{G} \in \mathbb{R}^{N_{\text{lat}} \times N_{\text{lon}} \times 2}$, respectively. The correction model learns the mapping

$$\hat{\mathbf{Y}} = f_{\theta}(\mathbf{X}, \mathbf{E}, \mathbf{S}, \mathbf{G}), \quad (2)$$

with learnable parameters θ . The observation sequence $\mathbf{X}$ ends at the forecast origin, whereas $\mathbf{E}$ covers valid times $t + 1$ to $t + T_{\text{out}}$; no future station measurement is available to the model. As illustrated in Figure 1, the task is therefore to condition a physically generated future NWP trajectory on recent station behavior and spatial context, rather than to extrapolate the future from observations alone.

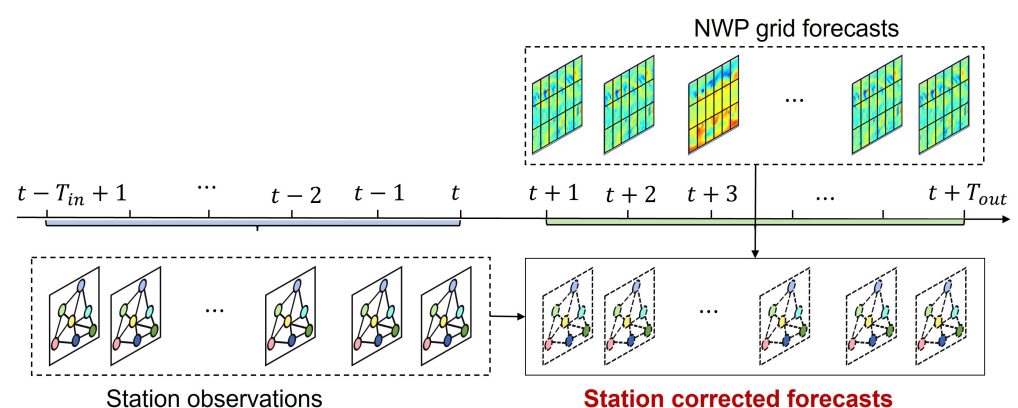


Figure 1. Input–output definition of station-level NWP bias correction. Historical station observations end at the forecast origin, while gridded NWP provides guidance over the future horizon. The target is the corresponding corrected trajectory at each station. The different colors in the NWP grid (top right) represent the magnitude of the target weather variable (such as temperature), with warmer colors indicating higher values.

2.2. TF-STNet Overview and Station-Centered NWP Extraction

TF-STNet addresses the source, spatial-support, and representation mismatches identified in the Introduction through four stages (Figure 2): station-centered NWP extraction, parallel time- and frequency-domain pathways, and prediction-level fusion. The extraction stage maps the regular NWP grid to a common local support around each irregular station. The time-domain pathway learns lead-specific corrections from recent observations, future NWP, and inter-station dependence, whereas the frequency-domain pathway targets horizon-scale distortions in oscillatory strength and timing. Their outputs are fused only after both pathways have produced station-level trajectories. The hidden dimension is denoted by D and the number of retained NWP cells per station by K , and superscripts t and f identify time- and frequency-domain quantities. Subscripts attached to MLP and Head distinguish independently parameterized modules by their roles.

The shared extraction stage converts the regional NWP lattice into a compact station-centered representation. For station $i \in \{1, \dots, N\}$ with coordinate vector $\mathbf{s}_i$, let $\mathcal{N}_K(i, k) = (a_{ik}, b_{ik})$ denote the latitude–longitude index pair of its k th-nearest NWP grid point, where $k \in \{1, \dots, K\}$, $a_{ik} \in \{1, \dots, N_{\text{lat}}\}$ and $b_{ik} \in \{1, \dots, N_{\text{lon}}\}$. The local NWP tensor is

$$\mathbf{E}^K = \text{KNN}(\mathbf{E}, \mathbf{S}, \mathbf{G}) \in \mathbb{R}^{T_{\text{out}} \times N \times K \times C}, \quad E_{\tau, i, k, c}^K = E_{\tau, a_{ik}, b_{ik}, c}, \quad (3)$$

where $\tau \in \{1, \dots, T_{\text{out}}\}$ indexes forecast lead and $c \in \{1, \dots, C\}$ indexes meteorological variable. Rather than committing to a single interpolated grid value, Equation (3) retains K surrounding NWP trajectories as a physically meaningful candidate set. Subsequent learned operations determine how these candidates contribute to each correction. Sharing $\mathbf{E}^K$ across both pathways ensures that their differences arise from time- and frequency-domain representations rather than unequal spatial inputs.

From this shared local context, the time-domain pathway produces a lead-resolved candidate $\hat{\mathbf{Y}}^t$ after source-aware temporal alignment and adaptive graph propagation. In parallel, the frequency-domain pathway learns spectral relevance, evaluates amplitude and phase evidence across neighboring cells, and produces $\hat{\mathbf{Y}}^f$. Because both candidates use the same station, variable, and lead-time coordinates, they can be combined directly without forcing their intermediate features into an artificial common latent space.

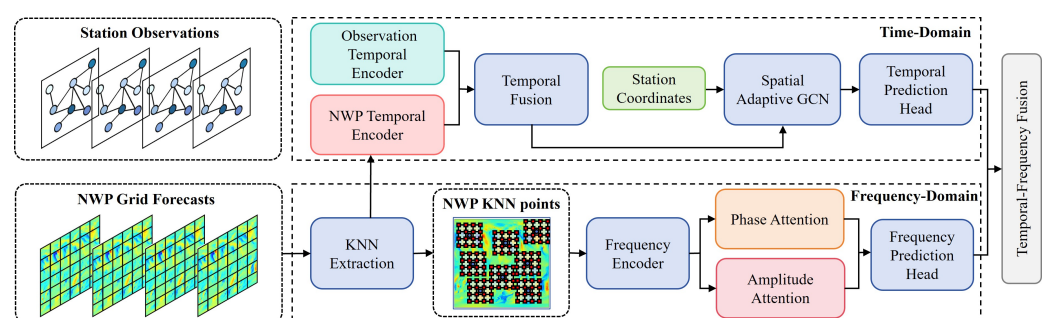


Figure 2. Architecture of TF-STNet. Shared KNN extraction supplies the same local NWP trajectories to two complementary pathways. Source-aware temporal alignment and adaptive graph propagation produce the time-domain candidate, while spectral weighting and amplitude–phase attention produce the frequency-domain candidate. Prediction-level fusion combines the two station-level trajectories. The different colors in the NWP grid (bottom left) represent the magnitude of the target weather variable (such as temperature), with warmer colors indicating higher values.

2.3. Time-Domain Pathway

2.3.1. Source-Aware Temporal Encoding

Historical observations and future NWP are processed by separate temporal encoders because they differ in temporal direction, spatial organization, and error characteristics.

Both encoders use gated dilated causal convolutions, but they do not share parameters: $\mathbf{X}$ is a measured history ending at t , whereas $\mathbf{E}^K$ is a model-generated future sequence with an additional local-neighbor axis. For a generic hidden sequence $\mathbf{U}$, a dilated causal convolution with kernel Ψ at temporal index τ is

$$(\mathbf{U} *_{\delta} \Psi)_{\tau} = \sum_{r=0}^{L_c-1} \mathbf{U}_{\tau-\delta r} \Psi_r, \quad (4)$$

where L_c is the kernel length, δ is the dilation factor, r indexes the kernel tap, and unavailable indices are replaced by left causal padding. Increasing δ across successive blocks expands the receptive field without discarding the temporal order.

Each block applies two convolutions to form a candidate state and a data-dependent gate:

$$\mathcal{T}_{\delta}(\mathbf{U}) = \tanh(\mathbf{U} *_{\delta} \Theta_a + \mathbf{b}_a) \odot \sigma(\mathbf{U} *_{\delta} \Theta_g + \mathbf{b}_g), \quad (5)$$

where Θ_a and Θ_g are the learnable convolution kernels for the candidate and gate, $\mathbf{b}_a$ and $\mathbf{b}_g$ are their biases, $\odot$ denotes element-wise multiplication, and $\sigma(\cdot)$ is the sigmoid function. Residual connections preserve short-range fluctuations that could otherwise be attenuated by the dilated stack.

The observation encoder produces $\mathbf{H}^{\text{obs}} \in \mathbb{R}^{T_{\text{in}} \times N \times D}$, while the NWP encoder produces $\mathbf{H}^{\text{nwp}} \in \mathbb{R}^{T_{\text{out}} \times N \times K \times D}$. Direct fusion is inappropriate because the former is indexed by historical time and the latter by forecast lead. A learned horizon projection $\mathcal{Q}$ maps the observation history to lead-conditioned station states, and $\mathcal{P}_K$ denotes a learnable aggregation along the K -neighbor axis. The aligned features are then fused as

$$\mathbf{Z}^t = \text{MLP}_t(\mathcal{Q}(\mathbf{H}^{\text{obs}}) \parallel \mathcal{P}_K(\mathbf{H}^{\text{nwp}})) \in \mathbb{R}^{T_{\text{out}} \times N \times D}. \quad (6)$$

Thus, $\mathbf{Z}^t$ represents how the future NWP trajectory should be corrected at each lead in light of the station's recent evolution. Observations condition the physically guided future trajectory rather than being used to extrapolate it independently.

2.3.2. Coordinate-Conditioned Adaptive Station Graph

The aligned temporal states are initially station-specific, although meteorological disturbances extend across multiple sites. TF-STNet therefore propagates them over an adaptive station graph. Coordinates $\mathbf{S}$ provide a stable geographic prior, while separate source and target embeddings allow learned dependence to depart from a symmetric distance kernel. This flexibility accommodates directional relationships induced by terrain, coastlines, and prevailing flow. The embeddings are

$$\mathbf{U}_{\text{src}} = \text{MLP}_{\text{src}}(\mathbf{S}), \quad \mathbf{U}_{\text{tgt}} = \text{MLP}_{\text{tgt}}(\mathbf{S}), \quad \mathbf{U}_{\text{src}}, \mathbf{U}_{\text{tgt}} \in \mathbb{R}^{N \times D_g}, \quad (7)$$

where D_g is the graph-embedding dimension. Their compatibility scores define a directed, row-normalized adaptive adjacency matrix,

$$\mathbf{A} = \text{Softmax}_{\text{row}}(\text{ReLU}(\mathbf{U}_{\text{src}} \mathbf{U}_{\text{tgt}}^{\top})) \in \mathbb{R}^{N \times N}. \quad (8)$$

Separate source and target embeddings permit $A_{ij} \neq A_{ji}$ for station indices i and j . At each forecast lead τ , the graph layer combines messages from related stations with a station-wise residual transformation:

$$\mathbf{H}_{\tau}^s = \phi(\mathbf{A} \mathbf{Z}_{\tau}^t \mathbf{W}_g + \mathbf{Z}_{\tau}^t \mathbf{W}_{\text{res}}), \quad (9)$$

where $\mathbf{H}^s \in \mathbb{R}^{T_{\text{out}} \times N \times D}$ denotes the spatially propagated hidden states; $\mathbf{W}_g, \mathbf{W}_{\text{res}} \in \mathbb{R}^{D \times D}$ are learnable graph-message and residual projections, respectively; and ϕ is a nonlinear activation. The residual term limits excessive graph smoothing and retains station-specific behavior. A temporal prediction head then produces the time-domain candidate

$$\hat{\mathbf{Y}}^t = \text{Head}_{\text{time}}(\mathbf{H}^s) \in \mathbb{R}^{T_{\text{out}} \times N \times C}. \quad (10)$$

2.4. Frequency-Domain Pathway

2.4.1. Spectral Encoding

The time-domain pathway emphasizes temporal order and lead-specific transitions but does not explicitly isolate oscillatory structure across the complete horizon. The frequency-domain pathway provides this complementary view using the same local NWP tensor $\mathbf{E}^K$, thereby remaining anchored to the physically generated future guidance. By reindexing the forecast leads as $\tau \in \{0, \dots, T_{\text{out}} - 1\}$ for the transform, the discrete Fourier coefficient for each station–neighbor–variable tuple (i, k, c) is

$$F_{\omega, i, k, c} = \sum_{\tau=0}^{T_{\text{out}}-1} E_{\tau, i, k, c}^K \exp\left(-j \frac{2\pi\omega\tau}{T_{\text{out}}}\right), \quad (11)$$

where ω is the frequency-bin index, $\Omega = \lfloor T_{\text{out}}/2 \rfloor + 1$ is the number of retained bins, and $j^2 = -1$. Thus, $\omega \in \{0, \dots, \Omega - 1\}$. Only the nonredundant coefficients of the real-valued trajectory are retained. A bounded spectral gate $\mathbf{w}_f \in \mathbb{R}^\Omega$ learns the relative relevance of these frequency bins:

$$\tilde{F}_{\omega, i, k, c} = F_{\omega, i, k, c} \sigma(w_{f, \omega}). \quad (12)$$

The gate is shared across stations and local neighbors within each regional model, encouraging region-level spectral selection rather than sample-limited station-specific filtering. The sigmoid bounds each multiplier, preventing unstable amplification while attenuating weakly informative frequency bins.

Collecting the coefficients into $\tilde{\mathbf{F}} \in \mathbb{C}^{\Omega \times N \times K \times C}$, the gated spectrum is decomposed into amplitude and phase:

$$\mathbf{M} = |\tilde{\mathbf{F}}|, \quad \Phi = \text{atan2}(\text{Im } \tilde{\mathbf{F}}, \text{Re } \tilde{\mathbf{F}}), \quad (13)$$

where $\mathbf{M}, \Phi \in \mathbb{R}^{\Omega \times N \times K \times C}$. Amplitude describes the strength of an oscillatory component, whereas phase locates it relative to the forecast origin. Treating them separately prevents a neighboring cell with realistic variability but incorrect timing from being considered equivalent to one with accurate timing but distorted variability. This distinction is important for fronts, sea-breeze cycles, and terrain-modulated responses whose timing can vary across nearby grid cells.

2.4.2. Amplitude–Phase Attention and Frequency-Domain Prediction

Amplitude and phase are projected independently into a common hidden dimension, allowing each component to learn a distinct nonlinear representation before neighborhood aggregation:

$$\mathbf{Z}^M = \text{MLP}_M(\mathbf{M}), \quad \mathbf{Z}^\Phi = \text{MLP}_\Phi(\Phi), \quad \mathbf{Z}^M, \mathbf{Z}^\Phi \in \mathbb{R}^{\Omega \times N \times K \times D}. \quad (14)$$

For $q \in \{M, \Phi\}$ and fixed (ω, i) , let $\mathbf{Z}_{\omega, i}^q \in \mathbb{R}^{K \times D}$ denote the hidden vectors of the K neighboring NWP cells. Scaled dot-product attention is then applied only over this local-neighbor axis:

$$\begin{aligned}\mathbf{Q}_{\omega,i}^q &= \mathbf{Z}_{\omega,i}^q \mathbf{W}_Q^q, & \mathbf{K}_{\omega,i}^q &= \mathbf{Z}_{\omega,i}^q \mathbf{W}_K^q, & \mathbf{V}_{\omega,i}^q &= \mathbf{Z}_{\omega,i}^q \mathbf{W}_V^q, \\ \mathbf{O}_{\omega,i}^q &= \text{Attn}_K(\mathbf{Z}_{\omega,i}^q)\end{aligned}\quad (15)$$

$$= \text{Softmax}_K \left(\frac{\mathbf{Q}_{\omega,i}^q (\mathbf{K}_{\omega,i}^q)^\top}{\sqrt{D}} \right) \mathbf{V}_{\omega,i}^q \quad (16)$$

where $\mathbf{W}_Q^q, \mathbf{W}_K^q, \mathbf{W}_V^q \in \mathbb{R}^{D \times D}$ are component-specific query, key, and value projections, respectively. For fixed (ω, i, q) , the resulting $\mathbf{Q}_{\omega,i}^q, \mathbf{K}_{\omega,i}^q, \mathbf{V}_{\omega,i}^q$, and $\mathbf{O}_{\omega,i}^q$ all belong to $\mathbb{R}^{K \times D}$, and Softmax_K normalizes each query over the K neighboring cells. Attention is therefore restricted to the local-neighbor axis and conditioned on frequency and station, so the preferred NWP cell may vary across locations and oscillatory scales. Consequently, $\mathbf{O}^M$ can favor cells that better represent variation strength, while $\mathbf{O}^\Phi$ can favor different cells with more consistent timing. Component-specific parameters preserve this distinction until the two forms of evidence are combined.

The two attention outputs are concatenated and mapped to a joint spectral representation:

$$\mathbf{H}^f = \mathcal{P}_K \left[\text{MLP}_f \left(\mathbf{O}^M \parallel \mathbf{O}^\Phi \right) \right]. \quad (17)$$

Here, the frequency-pathway instance of $\mathcal{P}_K$ aggregates the attended neighbor axis, yielding $\mathbf{H}^f \in \mathbb{R}^{\Omega \times N \times D}$. A frequency prediction head then performs learned spectral-to-horizon reconstruction and variable projection:

$$\hat{\mathbf{Y}}^f = \text{Head}_{\text{freq}}(\mathbf{H}^f) \in \mathbb{R}^{T_{\text{out}} \times N \times C}. \quad (18)$$

Because separate attention and nonlinear fusion mix evidence from multiple neighboring trajectories, $\mathbf{H}^f$ is not the spectrum of a single physical signal and cannot be reconstructed by a direct inverse Fourier transform. The learned head instead maps the selected spectral evidence to the lead-resolved frequency-domain candidate.

2.5. Prediction-Level Time–Frequency Fusion and Training Objective

The two pathway outputs are correction candidates with the same $T_{\text{out}} \times N \times C$ station-by-variable forecast structure. They are fused as

$$\hat{\mathbf{Y}} = \text{MLP}_o \left(\hat{\mathbf{Y}}^t \parallel \hat{\mathbf{Y}}^f \right). \quad (19)$$

Fusion at the prediction level combines tensors with the same physical interpretation and avoids imposing arbitrary alignment between lead-indexed temporal states and frequency-indexed features. It also preserves pathway specialization while MLP_o learns a nonlinear combination that can vary across variables, stations, and forecast leads. All modules, from local extraction to the final output head, are trained end to end using mean absolute error (MAE):

$$\mathcal{L}_{\text{MAE}} = \frac{1}{T_{\text{out}} N C} \sum_{\tau=1}^{T_{\text{out}}} \sum_{i=1}^N \sum_{c=1}^C \left| \hat{Y}_{\tau,i,c} - Y_{\tau,i,c} \right|. \quad (20)$$

MAE matches the primary evaluation metric and is less sensitive than squared loss to a small number of large residuals.

3. Experimental Design

3.1. Study Regions and Data

The evaluation uses hourly station observations and corresponding NWP forecasts from 1 May 2024 to 30 November 2025. The dataset covers 455 national meteorological

stations in four geographically distinct regions of China: Shandong (109 stations; 34–38° N, 115–123° E), Fujian (60 stations; 23–28° N, 116–121° E), Sichuan (145 stations; 26–34° N, 97–109° E), and Hebei (141 stations; 36–42° N, 113–120° E). The NWP fields have a horizontal resolution of 0.1° and are stored as regional tensors containing 65×68 , 85×42 , 50×50 , and 113×85 grid cells, respectively. As shown in Figure 3, these regions differ substantially in terrain and station density, providing diverse conditions for evaluating grid-to-station correction.

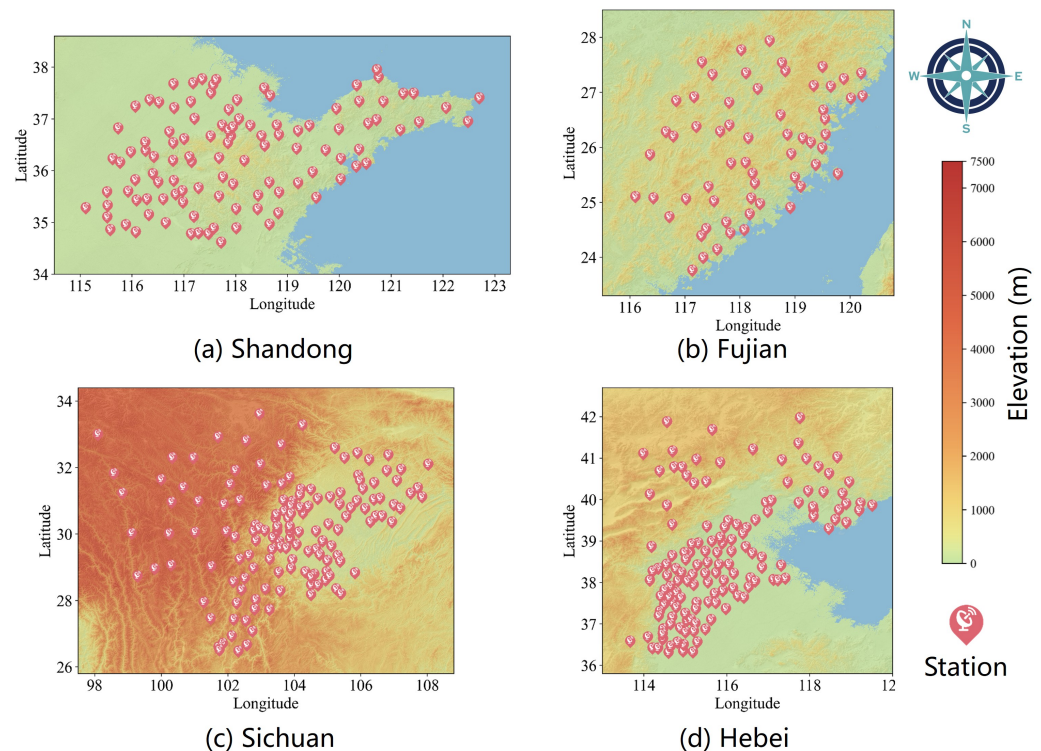


Figure 3. Terrain and station distributions in the four study regions: (a) Shandong, (b) Fujian, (c) Sichuan, and (d) Hebei. Background color denotes elevation, and markers indicate meteorological stations.

Table 1 compares the observed and raw NWP distributions. The discrepancies vary substantially across both variables and regions. Mean NWP wind speed exceeds the observed mean by 1.47, 1.21, and 1.16 m s^{-1} in Shandong, Fujian, and Hebei, respectively, but is nearly unbiased in Sichuan. Pressure exhibits a different pattern: the Sichuan observation and NWP means are 888 and 769 hPa, respectively, a 119 hPa discrepancy. This unusually large offset is consistent with the sensitivity of pressure to differences between grid-cell terrain height and station elevation; it helps explain why pressure shows the largest correction gains in this study and why raw pressure should not be compared across regions without considering elevation. In Hebei, raw NWP pressure is lower on average and substantially more variable than the observations. These additive, dispersion, and region-dependent errors motivate a conditional station-correction model rather than a uniform adjustment.

Table 1. Observed and raw NWP statistics in the four study regions (mean $\pm$ standard deviation).

Region	Source	Wind Speed (m s ⁻¹)	Pressure (hPa)	Relative Humidity (%)	Temperature (°C)
Shandong	Observation	2.36 $\pm$ 1.72	1004.00 $\pm$ 19.49	63.99 $\pm$ 24.00	17.70 $\pm$ 10.58
	NWP	3.83 $\pm$ 2.44	1008.00 $\pm$ 13.34	62.20 $\pm$ 22.78	17.70 $\pm$ 10.00
Fujian	Observation	2.14 $\pm$ 1.76	986.00 $\pm$ 29.60	75.90 $\pm$ 17.66	22.00 $\pm$ 7.57
	NWP	3.35 $\pm$ 3.54	976.00 $\pm$ 32.09	76.90 $\pm$ 17.35	21.50 $\pm$ 7.26
Sichuan	Observation	1.69 $\pm$ 1.30	888.00 $\pm$ 32.42	72.40 $\pm$ 21.24	17.80 $\pm$ 9.21
	NWP	1.68 $\pm$ 1.30	769.00 $\pm$ 43.19	70.75 $\pm$ 22.25	11.06 $\pm$ 10.44
Hebei	Observation	1.93 $\pm$ 1.57	993.00 $\pm$ 37.78	60.70 $\pm$ 25.46	16.00 $\pm$ 11.53
	NWP	3.09 $\pm$ 2.11	953.00 $\pm$ 63.65	56.70 $\pm$ 24.04	13.70 $\pm$ 11.87

3.2. Quality Control and Sample Construction

Observation and NWP timestamps, time zones, and physical units were first harmonized. Missing observation intervals of at most 3 h were filled by temporal interpolation, whereas windows containing longer gaps were excluded. Potentially implausible values were identified using variable-specific physical ranges and a z-score rule and then checked against adjacent times and neighboring stations. NWP forecasts were aligned with station observations by initialization time, lead time, and valid time. The raw element-wise missing rates recorded before this quality-control step were 3.0–3.2% in training, approximately 1.0% in validation, and 2.1–2.4% in testing (Table 2). Because the retained quality-control records provide element-wise rather than window-level counts, the exact fraction of windows removed by the long-gap rule cannot be reconstructed. The records also do not indicate whether an interpolated target within a retained window was used as a scored label. We therefore report this detail as unresolved rather than inferring an undocumented policy. At evaluation, any target position that remains unavailable is excluded from both the absolute-error sum and the denominator. With $v_m \in \{0, 1\}$ indicating whether target position m is available, the masked metric is $\text{MAE}_{\text{mask}} = \sum_m v_m |\hat{y}_m - y_m| / \sum_m v_m$.

Table 2. Raw observation missing rates before interpolation or long-gap exclusion. Rates are element-wise percentages pooled over the four variables; the within-region rates were reported as consistent across variables.

Region	Training	Validation	Test
Fujian	3.0%	1.0%	2.1%
Shandong	3.1%	1.0%	2.2%
Hebei	3.2%	1.0%	2.3–2.4%
Sichuan	3.1%	1.0%	2.4%

Samples were partitioned chronologically into 9752 training, 2098 validation, and 2098 test windows, corresponding to approximately 70%, 15%, and 15% of the data. Preserving temporal order prevents information from later periods from entering model training. This split tests temporal generalization within the four regional station networks; it does not test transfer to an unseen station, region, or climate regime. Consistent with the problem formulation, each sample contains 24 h of historical station observations, 24 h of future NWP guidance, and 24 h of future station targets. Normalization parameters were estimated from the training partition and applied unchanged to the validation and test partitions.

3.3. Baselines, Model Selection, and Evaluation Metrics

The evaluation compares TF-STNet with raw NWP and five learned baselines: Transformer [21], PatchTST [23], DLinear [24], Graph WaveNet (GWNNet) [17], and MegaCRN [18]. Transformer, PatchTST, and DLinear are representative sequence-modeling

baselines, whereas GWNet and MegaCRN are graph-based spatiotemporal forecasters. For a forecast issued at time t , all learned methods use station observations from $t - 23$ to t , future NWP guidance from $t + 1$ to $t + 24$, and station targets over the same $t + 1$ to $t + 24$ period. Every learned baseline is evaluated on the same chronological partitions and forecast cases. The sequence baselines do not use explicit coordinate information, whereas the graph-based baselines and TF-STNet incorporate geographic relationships through their respective graph representations. In TF-STNet, coordinate embeddings are used to construct the adaptive graph. The comparison therefore equalizes the principal data sources and temporal support while retaining architecture-specific spatial representations. Raw NWP is evaluated without any learned correction. GraphCast and related global foundation models are not treated as drop-in baselines because they require global, multilevel initial states and produce gridded atmospheric forecasts; replacing the supplied regional NWP would change the upstream forecast system rather than test NWP-to-station post-processing. Their omission is therefore a scope limitation, not a claim that they are inferior.

Each method was trained and evaluated with random seeds 2026, 2025, 2024, 2023, and 2022 using the same split and evaluation procedure; we report the arithmetic mean and standard deviation across runs. For TF-STNet, K and D were selected independently by one-factor-at-a-time validation sweeps: $K \in \{1, \dots, 10\}$ and $D \in \{16, 32, 64, 128, 256\}$. These ranges span a single nearest cell through a broad local neighborhood and cover small to moderately overparameterized hidden representations. The task-specific selections are reported in Section 4.7.

Table 3 summarizes the fixed optimization, checkpoint, initialization, and seed settings. The public repository at https://github.com/gyla1993/TS-STNet_source_code (accessed on 3 September 2026) provides the model implementation, configuration files, and a wind-speed training and inference example.

Table 3. Fixed training configuration. Settings were held constant across regions, variables, and experimental analyses unless otherwise noted.

Item	Setting
Optimizer	Adam
Batch size	32
Training duration	300 epochs
Early stopping	Not used
Checkpoint rule	Retain the checkpoint with the lowest validation MAE after the completed 300-epoch run
Initial learning rate	0.001
Learning-rate schedule	None
Weight decay	0.0001
Dropout	0.1
Frequency-module initialization	Xavier uniform for linear weights; zero for linear biases; one for learnable frequency weights
Other-layer initialization	PyTorch 2.3.1 defaults
Reference/auxiliary seed	2026
Robustness seeds	2026, 2025, 2024, 2023, and 2022
Variable task selection	K and D selected independently by the lowest validation MAE; all other settings fixed

Table 4 summarizes the input availability and temporal alignment for the compared methods.

The primary evaluation metric is MAE:

$$\text{MAE} = \frac{1}{M} \sum_{m=1}^M |\hat{y}_m - y_m|, \quad (21)$$

where M is the total number of scalar predictions across all test windows, forecast leads, and stations within one region–variable task; m indexes those predictions, and $\hat{y}_m$ and y_m

are the corresponding prediction and observation. For the multi-seed comparison, MAE is complemented by the Pearson correlation coefficient (PCC), which measures the linear association between predictions and observations:

$$\text{PCC} = \frac{\sum_m (\hat{y}_m - \bar{\hat{y}})(y_m - \bar{y})}{\sqrt{\sum_m (\hat{y}_m - \bar{\hat{y}})^2} \sqrt{\sum_m (y_m - \bar{y})^2}}, \quad (22)$$

where $\bar{\hat{y}}$ and $\bar{y}$ are the means of predictions and observations over the same valid positions. Because MAE has different units and scales across variables, aggregate performance for method q in region r is summarized by normalized MAE (NMAE):

$$\text{NMAE}_{q,r} = \frac{1}{C} \sum_{c=1}^C \frac{\text{MAE}_{q,r,c}}{\text{MAE}_{\text{NWP},r,c}}. \quad (23)$$

Here, q indexes the forecasting method, r the region, and c the meteorological variable; $\text{MAE}_{\text{NWP},r,c}$ is the raw NWP MAE for region r and variable c . Raw NWP therefore has an NMAE of 1 by definition, while lower values indicate larger aggregate correction gains. The four-region mean NMAE reported in the Results averages Equation (23) across the four study regions.

Table 4. Information available to each method in the reported comparison. A checkmark (✓) indicates that the corresponding input is used or, in the target column, that forecasts are evaluated over the indicated target horizon. Blank cells indicate inputs that are not used.

Method	Station History $t - 23:t$	Future NWP $t + 1:t + 24$	Coordinate Information	Target $t + 1:t + 24$
Raw NWP		✓		✓
Transformer/PatchTST/DLinear	✓	✓		✓
GWNet/MegaCRN	✓	✓	✓	✓
TF-STNet	✓	✓	✓	✓

4. Results and Discussion

4.1. Improvement over Raw NWP

This analysis assesses whether TF-STNet consistently corrects raw NWP across meteorological variables and geographic regions, rather than deriving its overall gain from a small number of favorable tasks. We report the five-run mean and standard deviation for each method and examine the percentage reduction in MAE for each of the 16 region–variable combinations.

Table 5 provides the updated five-run results. TF-STNet has the lowest mean MAE in 15 of the 16 tasks and the highest mean PCC in 15 tasks. The only MAE exception is Shandong–temperature (0.485 ± 0.017), for which MegaCRN (0.430 ± 0.082) has lower errors. The pressure MAE reductions relative to the best learned comparator are 16.2% in Shandong, 26.6% in Fujian, 17.0% in Sichuan, and 57.4% in Hebei. Standard deviations are small for TF-STNet in most tasks, whereas several baselines show larger variation in pressure and temperature.

In summary, the multi-seed analysis supports broad but not universal superiority over the learned baselines. The raw NWP reduction remains strongly variable-dependent, with pressure showing the largest gains because its grid-to-station offset is especially systematic; temperature, particularly in Shandong, is less uniformly correctable.

Table 5. Five-seed test-set comparison with the best learned comparator by mean MAE for each task. Values are mean $\pm$ standard deviation over seeds 2026, 2025, 2024, 2023, and 2022. PCC is Pearson correlation; a negative reduction indicates that the proposed model has higher MAE.

Region	Variable	TF-STNet		Method	Best Comparator		MAE Reduction vs. Comparator
		MAE	PCC		MAE	PCC	
Shandong	Wind	0.676 $\pm$ 0.001	0.700 $\pm$ 0.000	MegaCRN	0.702 $\pm$ 0.003	0.677 $\pm$ 0.001	3.7%
	Pressure	0.353 $\pm$ 0.144	0.998 $\pm$ 0.002	Transformer	0.421 $\pm$ 0.068	0.997 $\pm$ 0.000	16.2%
	RHU	7.000 $\pm$ 0.374	0.855 $\pm$ 0.013	DLinear	7.865 $\pm$ 1.054	0.801 $\pm$ 0.055	11.0%
Fujian	Temperature	0.485 $\pm$ 0.017	0.996 $\pm$ 0.000	MegaCRN	0.430 $\pm$ 0.082	0.997 $\pm$ 0.001	−12.8%
	Wind	0.709 $\pm$ 0.005	0.634 $\pm$ 0.003	GWNet	0.739 $\pm$ 0.007	0.602 $\pm$ 0.009	4.1%
	Pressure	0.405 $\pm$ 0.010	0.990 $\pm$ 0.000	MegaCRN	0.552 $\pm$ 0.159	0.980 $\pm$ 0.011	26.6%
	RHU	5.003 $\pm$ 0.044	0.893 $\pm$ 0.001	GWNet	5.313 $\pm$ 0.263	0.861 $\pm$ 0.019	5.8%
Sichuan	Temperature	0.992 $\pm$ 0.016	0.960 $\pm$ 0.001	GWNet	1.067 $\pm$ 0.100	0.940 $\pm$ 0.012	7.0%
	Wind	0.640 $\pm$ 0.003	0.524 $\pm$ 0.003	Transformer	0.645 $\pm$ 0.001	0.501 $\pm$ 0.001	0.8%
	Pressure	1.082 $\pm$ 0.077	0.954 $\pm$ 0.005	MegaCRN	1.303 $\pm$ 0.255	0.902 $\pm$ 0.035	17.0%
	RHU	6.131 $\pm$ 0.080	0.872 $\pm$ 0.002	MegaCRN	6.222 $\pm$ 0.067	0.856 $\pm$ 0.006	1.5%
Hebei	Temperature	1.124 $\pm$ 0.018	0.958 $\pm$ 0.001	Transformer	1.279 $\pm$ 0.081	0.945 $\pm$ 0.006	12.1%
	Wind	0.606 $\pm$ 0.005	0.688 $\pm$ 0.001	GWNet	0.621 $\pm$ 0.004	0.658 $\pm$ 0.002	2.4%
	Pressure	0.467 $\pm$ 0.026	0.997 $\pm$ 0.000	GWNet	1.095 $\pm$ 0.186	0.984 $\pm$ 0.005	57.4%
	RHU	7.487 $\pm$ 0.235	0.870 $\pm$ 0.003	GWNet	7.690 $\pm$ 0.017	0.849 $\pm$ 0.004	2.6%
	Temperature	1.078 $\pm$ 0.017	0.983 $\pm$ 0.001	Transformer	1.147 $\pm$ 0.247	0.969 $\pm$ 0.010	6.0%

Figure 4 visualizes the corresponding task-wise MAE reductions relative to raw NWP, making the strong pressure improvements and the greater regional variation in temperature readily apparent.

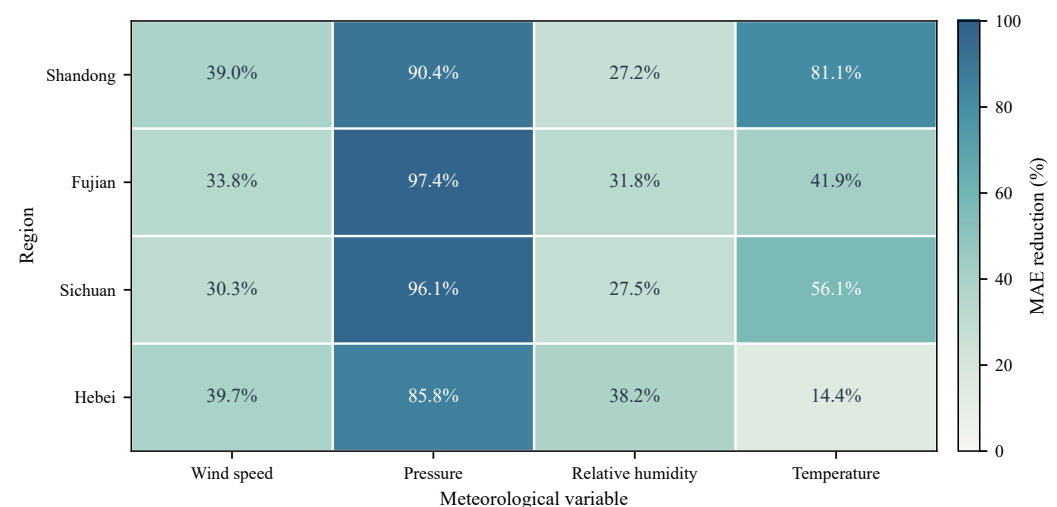


Figure 4. Percentage reduction in five-seed mean MAE achieved by TF-STNet relative to raw NWP for each region–variable task. Higher values indicate greater improvement; negative values would indicate degradation.

4.2. Comparison with Forecasting Baselines

Table 6 summarizes NMAE computed from the five-run mean MAE values. TF-STNet achieves the lowest NMAE in every region and a four-region mean of 0.481. After correction of the Shandong–temperature entry, GWNet is the strongest aggregate comparator, with a mean of 0.531, so the relative reduction is 9.4%; MegaCRN has a corrected four-region mean of 0.558. The regional NMAE reductions relative to MegaCRN are 10.7% in Shandong, 9.1% in Fujian, 5.4% in Sichuan, and 24.2% in Hebei. These aggregate values should be read together with the task-wise standard deviations in Table 5; NMAE is a descriptive scale-normalized summary, not an additional independent test.

Although the NMAE of TF-STNet ranges from 0.406 in Shandong to 0.555 in Hebei, it ranks first in each region. Thus, the aggregate improvement is observed across all four study areas rather than being driven by a single region.

Table 6. Regional NMAE of the compared methods and the four-region mean. Lower is better.

Method	Shandong	Fujian	Sichuan	Hebei	Mean
NWP	1.000	1.000	1.000	1.000	1.000
Transformer	0.521	0.557	0.524	0.658	0.565
PatchTST	0.624	0.733	0.678	0.834	0.717
DLinear	0.507	0.551	0.543	0.733	0.584
GWNet	0.460	0.522	0.511	0.629	0.531
MegaCRN	0.455	0.537	0.502	0.732	0.558
TF-STNet	0.406	0.488	0.475	0.555	0.481

The four-region mean NMAE table provides an aggregate comparison of the five-run means. TF-STNet has the lowest four-region mean MAE for all four variables, although Table 5 shows that this aggregate result does not imply the lowest MAE in every individual task. In particular, Shandong–temperature remains better served by several baselines. Thus, the aggregate advantage is not driven solely by the highly correctable pressure field, but the temperature result also cautions against claiming universal task-wise dominance.

GWNet and MegaCRN generally outperform the sequence-only baselines, a pattern consistent with the value of explicit spatial modeling. PatchTST is less stable in pressure and temperature, as indicated by its larger five-run deviations. The further improvement achieved by TF-STNet, together with the frequency-pathway ablation below, suggests that adaptive station graphs do not capture all useful structure in local NWP trajectories and that complementary spectral modeling provides additional information. Figure 5 complements the regional NMAE summary by comparing the four-region mean MAE of all methods separately for each meteorological variable.

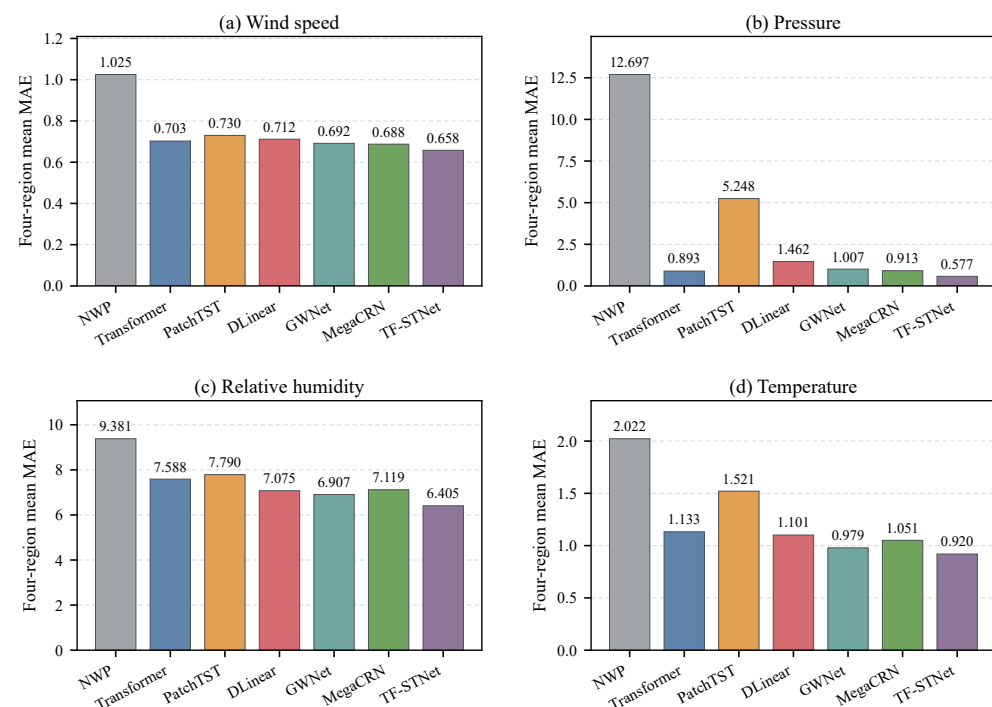


Figure 5. Four-region mean MAE of raw NWP and the learned forecasting methods, computed from the five-seed means, for (a) wind speed, (b) pressure, (c) relative humidity, and (d) temperature. Lower values indicate better performance.

4.3. Simple Pressure MOS Baselines

The large pressure correction gains could otherwise be interpreted as removal of an almost constant offset. We therefore compare TF-STNet with two deliberately simple model-output-statistics (MOS) baselines: mean-bias correction, $\hat{y} = x + b$ with $b = \text{mean}(y_{\text{train}} - x_{\text{train}})$, and univariate linear MOS, $\hat{y} = ax + b$ fitted on the training set. Both methods use only the local NWP pressure value and are evaluated on the same test cases as TF-STNet. This auxiliary analysis uses the reference seed 2026. Table 7 shows that TF-STNet has lower MAE in all four regions. The reduction relative to linear MOS is 27.3–84.3%, and the reduction relative to mean-bias correction is 44.3–86.5%. Thus, subtracting a single climatological bias does not explain the full pressure improvement; the result is consistent with additional contributions from station history, local spatial context, and the nonlinear lead-dependent mapping.

Table 7. Pressure MAE for simple MOS corrections and TF-STNet.

Region	Mean-Bias MOS	Linear MOS	TF-STNet	Reduction vs. Mean Bias	Reduction vs. Linear
Shandong	3.613	3.103	0.487	86.5%	84.3%
Fujian	1.804	1.382	0.405	77.5%	70.7%
Sichuan	1.943	1.488	1.082	44.3%	27.3%
Hebei	2.538	2.494	0.467	81.6%	81.3%

4.4. Forecast-Lead and Extreme-Event Evaluation

Aggregate MAE can conceal difficult leads or tail events. We therefore grouped the reference-seed test predictions by lead from 1 to 24 h and computed MAE separately at each lead. Figure 6 visualizes all 24 lead-specific errors, while Table 8 reports their mean and the leads attaining the minimum and maximum. The errors are non-monotonic: peaks occur around 6–11 h for several wind-speed, relative-humidity, and temperature series, while Shandong–temperature and Hebei–pressure increase near the end of the horizon. The model produces predictions throughout the 24-h horizon, but the lead-specific peaks identify the forecast periods with the largest errors. Because comparator outputs were not retained by lead, this analysis characterizes TF-STNet but does not establish a lead-wise advantage over the baselines.

Table 8. Lead-time error summary for TF-STNet. Each cell gives mean MAE across leads, followed by the minimum-lead and maximum-lead errors (hours).

Region	Wind	Pressure	RHU	Temperature
Shandong	0.671 (15/6)	0.474 (14/8)	6.367 (22/8)	0.614 (4/24)
Fujian	0.707 (19/8)	0.411 (16/8)	5.088 (22/9)	0.917 (20/9)
Sichuan	0.639 (24/8)	0.730 (17/1)	6.116 (23/10)	1.130 (19/9)
Hebei	0.598 (20/7)	0.441 (13/24)	7.401 (1/11)	1.035 (18/8)

Numbers in parentheses are the lead with the minimum MAE/the lead with the maximum MAE.

For tail verification, a high-wind event is a test position at which wind speed exceeds the training-set 90th percentile; a rapid-change event is a position at which the absolute change from the adjacent time exceeds the corresponding training-set 90th percentile. The thresholds are determined from training data only, and raw NWP, MegaCRN, and TF-STNet are evaluated on identical event sets using the reference seed 2026 for the learned models. Table 9 shows that TF-STNet has lower MAE than MegaCRN in all eight cases and than raw NWP in seven. The sole exception is the Hebei–high-wind event, for which the raw NWP and TF-STNet MAE are 1.427 and 1.523, respectively, although TF-STNet remains more

accurate than MegaCRN (1.687). Across the eight cases, the average improvements are 17.5% relative to raw NWP and 6.2% relative to MegaCRN. These results indicate improved performance for the tested wind events without implying universal superiority under rare conditions.

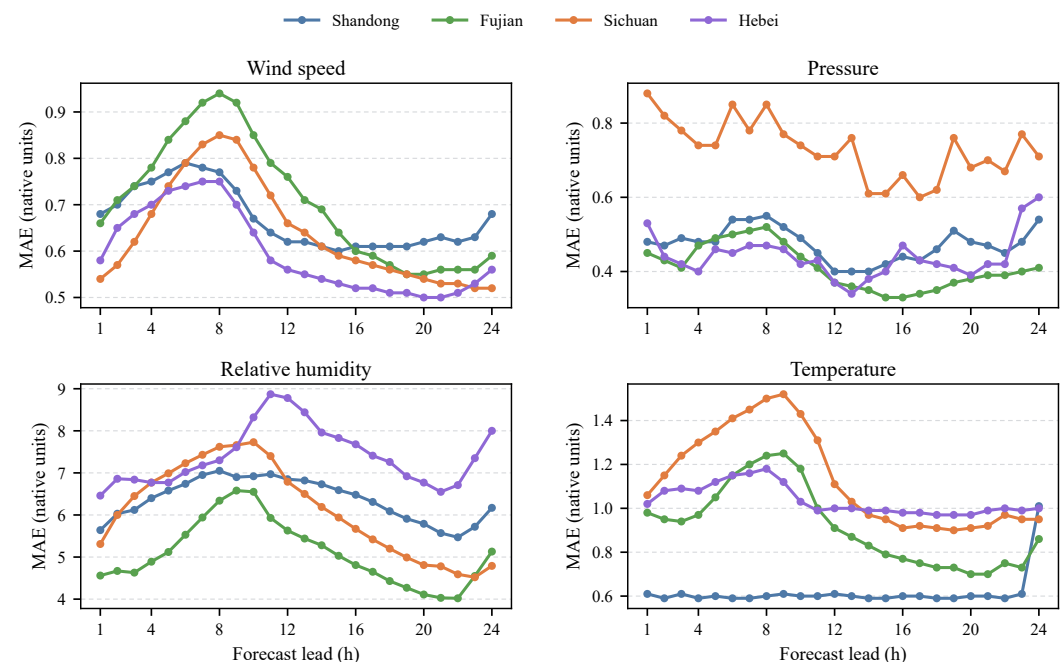


Figure 6. TF-STNet test-set MAE at each forecast lead from 1 to 24 h. Each panel is a meteorological variable, and each curve is one region; the independent y-scales preserve visibility across variables with different units and error magnitudes.

Table 9. High-wind and rapid-change events. Improvements are relative MAE reductions; positive values favor TF-STNet.

Region	Event	Samples	NWP	MegaCRN	TF-STNet	vs. NWP	vs. Mega
Fujian	High wind	11,555	2.766	1.849	1.808	34.6%	2.2%
Fujian	Rapid change	12,511	1.881	1.461	1.447	23.1%	1.0%
Shandong	High wind	9920	2.026	1.749	1.613	20.4%	7.8%
Shandong	Rapid change	14,864	1.602	1.370	1.322	17.5%	3.5%
Hebei	High wind	9443	1.427	1.687	1.523	−6.7%	9.7%
Hebei	Rapid change	15,605	1.447	1.308	1.257	13.1%	3.9%
Sichuan	High wind	23,992	2.491	2.308	2.008	19.4%	13.0%
Sichuan	Rapid change	26,232	1.725	1.534	1.410	18.3%	8.1%

4.5. Computational Efficiency and Reproducibility

Efficiency was measured on the same NVIDIA GeForce RTX 3080 Ti GPU with batch size 32. Table 10 reports wall-clock time for one training batch, one inference batch, and the number of trainable parameters. TF-STNet is consistently faster than MegaCRN (29.4–54.4% lower training time and 25.0–65.8% lower inference time) and has a similar though not identical cost to GWNet. It uses more parameters (0.837–0.941 million) than the two graph baselines, so the timing advantage is due to the measured implementation rather than a smaller model. The complete optimization, initialization, checkpoint, and seed settings are reported in Section 3.3. The public repository provides the model implementation, configuration files, and a wind-speed training and inference example.

Table 10. Computational cost on an NVIDIA RTX 3080 Ti (batch size 32). Times are seconds per batch.

Region	Model	Train/Batch (s)	Inference/Batch (s)	Params (M)
Shandong	GWNet	0.1564	0.0462	0.4943
	MegaCRN	0.2750	0.0923	0.0296
	TF-STNet	0.1253	0.0482	0.8863
Fujian	GWNet	0.0959	0.0273	0.4923
	MegaCRN	0.2051	0.0767	0.0276
	TF-STNet	0.0961	0.0262	0.8365
Sichuan	GWNet	0.1737	0.1031	0.4957
	MegaCRN	0.2997	0.1590	0.0310
	TF-STNet	0.2116	0.1192	0.9413
Hebei	GWNet	0.1317	0.0622	0.4955
	MegaCRN	0.2456	0.1202	0.0309
	TF-STNet	0.1590	0.0664	0.9344

4.6. Ablation Study

Four ablations separate the effects of input sources from those of the two representation pathways. The w/o NWP and w/o Obs. variants remove future NWP and historical observations, respectively; w/o Time and w/o Freq. remove the corresponding pathway while retaining the remaining inputs where applicable. Table 11 reports relative MAE, computed for each region by averaging the four ratios between an ablation's variable-wise MAE and that of the complete model. TF-STNet is normalized to 1.0, so values above 1 indicate degradation. The ablation analysis uses the reference seed 2026, whereas the main comparison in Table 5 reports the five-seed robustness analysis.

Removing the time-domain pathway causes the largest mean degradation, increasing relative MAE to 1.885, followed by removal of NWP at 1.819. The latter nearly doubles the error because it removes guidance over the future horizon, consistent with the intended role of TF-STNet as an NWP post-processor rather than an autoregressive station forecaster. The Sichuan w/o-Time value (2.899) is neither a percentage nor an absolute MAE; it is the mean of four variable-wise ratios. It is amplified by the pressure component, for which losing lead-resolved temporal states leaves a large terrain-related level error relative to the small complete-model MAE. Removing the frequency-domain pathway raises the mean relative MAE to 1.396, indicating that spectral amplitude and phase provide information not recovered by the time-domain pathway. Removing observations produces the smallest but still consistent degradation, to 1.154, quantifying the value of recent measurements for local state calibration. Together, these results indicate that future NWP and the time-domain pathway provide the largest aggregate contributions, while observations and frequency-domain learning provide complementary gains.

Table 11. Relative MAE of the ablation variants by region. Values are normalized to complete TF-STNet (1.000); values above 1 indicate degradation.

Variant	Shandong	Fujian	Sichuan	Hebei	Mean
w/o NWP	1.970	1.550	1.662	2.095	1.819
w/o Obs.	1.303	1.078	1.098	1.139	1.154
w/o Time	1.807	1.345	2.899	1.487	1.885
w/o Freq.	1.759	1.247	1.308	1.269	1.396
TF-STNet	1.000	1.000	1.000	1.000	1.000

Figure 7 presents the corresponding four-region mean MAE by variable. Complete TF-STNet achieves the lowest value in every panel. Without NWP, mean MAE increases to 0.767, 1.694, 7.779, and 1.393 for wind speed, pressure, relative humidity, and temperature, respectively. Removing the time-domain pathway raises pressure MAE to 2.303, 4.48 times

that of the complete model, while removing the frequency-domain pathway raises pressure and temperature MAE to 1.139 and 1.148. Because the figure reports cross-region means, it characterizes average pathway complementarity; Table 11 provides the corresponding region-specific behavior.

Pressure provides the clearest evidence of pathway complementarity: removing either pathway markedly increases its mean MAE, so the improvement cannot be attributed to adaptive graph propagation or spectral processing alone. The frequency-domain pathway also contributes substantially to temperature, for which its removal increases the mean MAE by 24.2%. Its smaller effect on wind speed and relative humidity may reflect their more intermittent, less strongly periodic behavior over a 24-h horizon. In contrast, pressure and temperature contain stronger low-frequency or diurnal structure, giving amplitude–phase separation more opportunity to distinguish systematic distortions. This explanation is an interpretation of the observed ablation pattern, not a causal claim about atmospheric dynamics. Although the regional penalties vary in magnitude, every ablation performs worse than the complete TF-STNet in every region, indicating that the pathway benefits are not artifacts of cross-region averaging.

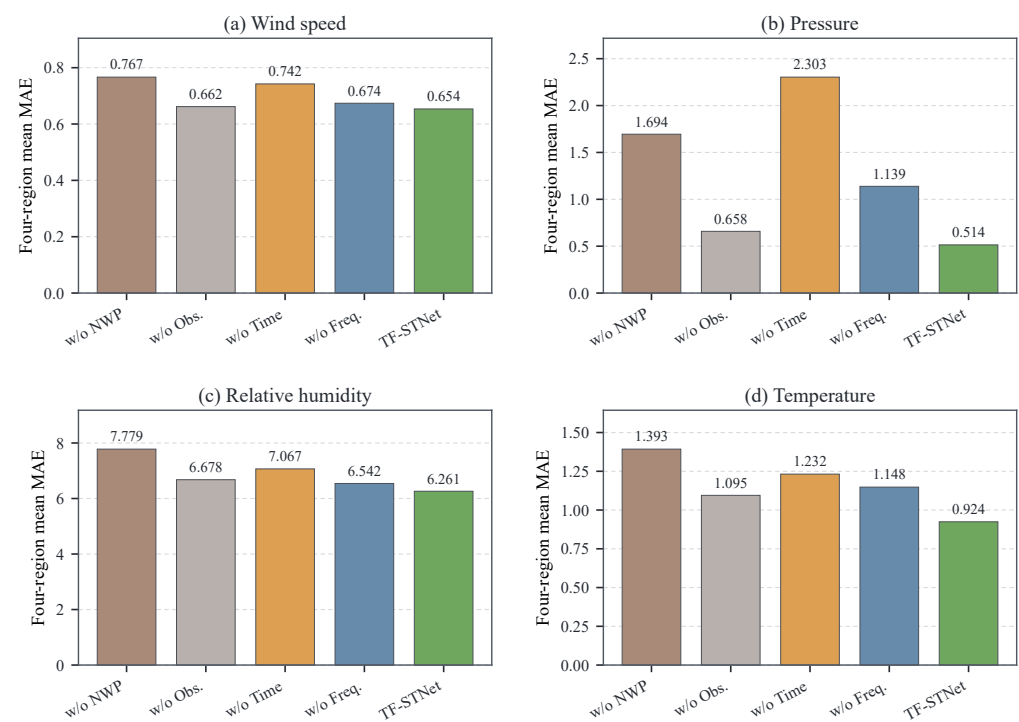


Figure 7. Four-region mean MAE of the ablation variants for (a) wind speed, (b) pressure, (c) relative humidity, and (d) temperature. Lower values indicate better performance.

4.7. Hyperparameter Sensitivity

Figure 8 consolidates the one-factor-at-a-time validation sweeps for all four variables. The top row varies the number of retained NWP neighbors K from 1 to 10, and the bottom row varies the hidden dimension D over 16, 32, 64, 128, and 256. Curves are plotted separately for each region because the error scales differ across variables; comparisons should therefore be made within a panel rather than between panels.

The response to K is non-monotonic and variable-dependent. For wind speed, the minima occur at $K = 5, 4, 7$, and 6 for Shandong, Fujian, Sichuan, and Hebei, respectively. The pressure optima shift to $K = 10, 3, 7$, and 7 ; temperature favors $K = 9, 2, 3$, and 2 ; and relative humidity favors $K = 2, 8, 4$, and 6 . This pattern indicates that a larger neighborhood is not uniformly beneficial: too few cells omit local context, whereas too many can introduce

weakly related trajectories. The preferred D values are similarly heterogeneous. Wind speed selects $D = 128, 128, 64$, and 256 across the four regions, while pressure selects $64, 32$, and 64 ; temperature selects $256, 128, 64$, and 64 ; and relative humidity selects $256, 256, 128$, and 128 . The $D = 16$ candidate is never optimal, but increasing capacity beyond the selected value does not yield monotonic gains. Table 12 lists every selected configuration and validation MAE. Together, Figure 8 and Table 12 show why K and D must be selected per region and variable rather than fixed globally.

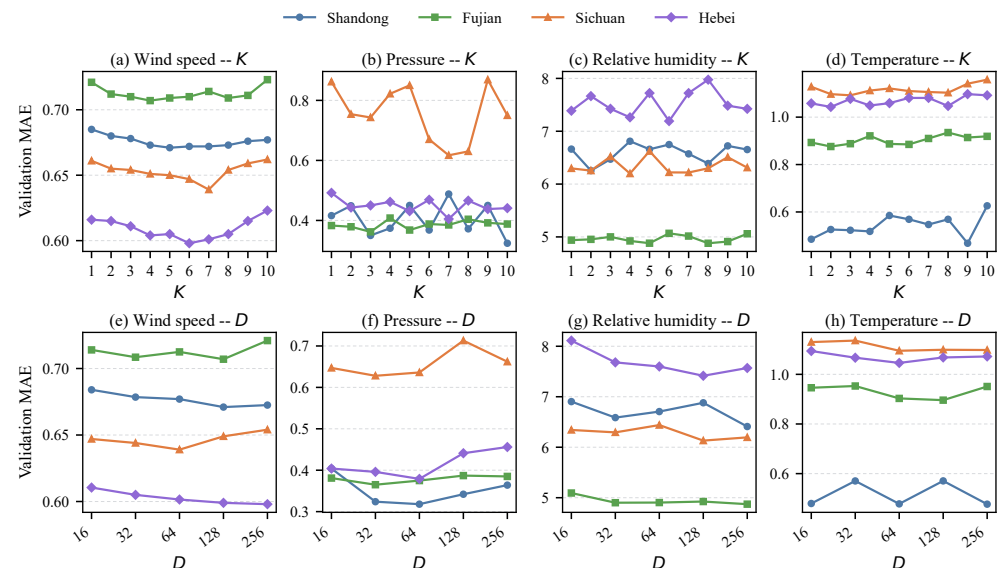


Figure 8. Unified validation-MAE sensitivity curves for all four meteorological variables. The top row varies the number of local NWP neighbors K and the bottom row varies the hidden dimension D ; each curve corresponds to one region. Lower values indicate better performance within a panel.

Table 12. Variable-specific hyperparameter selections from one-factor-at-a-time validation sweeps. The selected value and corresponding validation MAE are reported in separate columns for each sweep.

Region	Variable	Neighborhood Sweep		Dimension Sweep	
		Best K	Validation MAE	Best D	Validation MAE
Shandong	Wind speed	5	0.671	128	0.671
	Pressure	10	0.324	64	0.318
	Relative humidity	2	6.255	256	6.410
	Temperature	9	0.468	256	0.477
Fujian	Wind speed	4	0.707	128	0.707
	Pressure	3	0.362	32	0.365
	Relative humidity	8	4.880	256	4.872
	Temperature	2	0.876	128	0.896
Sichuan	Wind speed	7	0.639	64	0.639
	Pressure	7	0.617	32	0.628
	Relative humidity	4	6.199	128	6.131
	Temperature	3	1.092	64	1.095
Hebei	Wind speed	6	0.598	256	0.598
	Pressure	7	0.405	64	0.379
	Relative humidity	6	7.192	128	7.415
	Temperature	2	1.043	64	1.046

5. Conclusions

This study presents TF-STNet for translating gridded NWP forecasts into station-level meteorological trajectories. Its source-aware time-domain pathway aligns recent observations with future NWP and propagates lead-resolved states over an adaptive station graph. Its frequency-domain pathway evaluates local NWP neighborhoods through spectral

weighting and separate amplitude and phase attention. Prediction-level fusion combines the two candidate forecasts while preserving the distinct role of each representation.

Across five independent runs and 16 region-variable tasks spanning 455 stations in Hebei, Shandong, Fujian, and Sichuan, TF-STNet has the lowest mean MAE in 15 tasks and the highest mean PCC in 15 tasks. It achieves a four-region mean NMAE of 0.481, has lower pressure MAE than the simple MOS corrections in all four regions, and has lower wind-event MAE than MegaCRN in all eight tested cases. The ablation results indicate that future NWP, recent observations, and both pathways provide complementary information, with future NWP and the time-domain pathway making the largest aggregate contributions.

The evidence also defines the boundaries of these conclusions. The chronological split assesses temporal generalization within the four existing station networks, not transfer to unseen stations or regions. The evaluation is deterministic, so probabilistic calibration is not established. Event tests cover only high wind and rapid wind changes; moreover, the Hebei-high-wind case and the Shandong-temperature task show that the method is not uniformly the best. The available quality-control summary does not provide the exact fraction of windows excluded by the long-gap rule or establish whether interpolated targets were retained for scoring. Future work should consider probabilistic outputs, transfer to unseen stations and regions, broader event-stratified verification, and comparisons across alternative upstream forecast systems.

Author Contributions: Conceptualization, Z.W., B.X. and Y.G.; Methodology, Z.W. and B.X.; Software, Z.W., S.C. and Y.Y.; Validation, S.C., Y.Y. and B.W.; Formal analysis, Z.W. and S.C.; Investigation, Z.W., S.C., Y.Y. and B.W.; Resources, B.W., B.X. and Y.G.; Data curation, S.C., Y.Y. and B.W.; Writing—original draft, Z.W. and B.X.; Writing—review and editing, Z.W., S.C., Y.Y., B.W., B.X. and Y.G.; Visualization, Z.W. and S.C.; Supervision, B.X. and Y.G.; Project administration, B.W. and Y.G.; Funding acquisition, Z.W. and Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research study was funded by the Science and Technology Project of SGCC, grant number 5700-202455362A-3-1-DG.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study were provided by the State Key Laboratory of Renewable Energy Grid-Integration, China Electric Power Research Institute. The data are not publicly available but may be obtained from the corresponding author upon reasonable request, subject to approval by the data provider. The complete experimental configuration is reported in Section 3.3. The model implementation, configuration files, and a wind-speed training and inference example are available at https://github.com/gyla1993/TS-STNet_source_code (accessed on 3 September 2026).

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI Codex (CLI version 0.144.4) to polish the text and correct grammatical errors. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Authors Zhao Wang, Shuai Chen, Yunbo Yang, and Bo Wang were employed by China Electric Power Research Institute, Beijing, China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NWP	Numerical weather prediction
KNN	K-nearest-neighbor operator

TF-STNet	Time–frequency dual-branch spatiotemporal network
GWNet	Graph WaveNet
MegaCRN	Meta-graph convolutional recurrent network
GCN	Graph convolutional network
MOS	Model output statistics
MAE	Mean absolute error
PCC	Pearson correlation coefficient
NMAE	Normalized mean absolute error
RHU	Relative humidity
TMP	Temperature
MLP	Multilayer perceptron

References

- Xu, P.; Zheng, X.; Gao, T.; Wang, Y.; Yin, J.; Zhang, J.; Zhang, X.; Luo, S.; Wang, Z.; Zhang, Z.; et al. An Artificial Intelligence-Based Limited Area Model for Forecasting of Surface Meteorological Variables. *Commun. Earth Environ.* **2025**, *6*, 372. [\[CrossRef\]](#)
- Bodnar, C.; Bruinsma, W.P.; Lucic, A.; Stanley, M.; Allen, A.; Brandstetter, J.; Garvan, P.; Riechert, M.; Weyn, J.A.; Dong, H.; et al. A Foundation Model for the Earth System. *Nature* **2025**, *641*, 1180–1187. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tomasi, E.; Franch, G.; Cristoforetti, M. Can AI Be Enabled to Perform Dynamical Downscaling? A Latent Diffusion Model to Mimic Kilometer-Scale COSMO5.0_CLM9 Simulations. *Geosci. Model Dev.* **2025**, *18*, 2051–2078. [\[CrossRef\]](#)
- Rasp, S.; Lerch, S. Neural Networks for Postprocessing Ensemble Weather Forecasts. *Mon. Weather. Rev.* **2018**, *146*, 3885–3900. [\[CrossRef\]](#)
- Grönquist, P.; Yao, C.; Ben-Nun, T.; Dryden, N.; Dueben, P.; Li, S.; Hoefler, T. Deep Learning for Post-Processing Ensemble Weather Forecasts. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200092. [\[CrossRef\]](#) [\[PubMed\]](#)
- Huk, D.; Adewoyin, R.A.; Dutta, R. Probabilistic Rainfall Downscaling: Joint Generalized Neural Models with Censored Spatial Gaussian Copula. *J. Mach. Learn. Res.* **2026**, *27*, 1–46.
- Schmid, L.; Roidl, M.; Kirchheim, A.; Pauly, M. Comparing Statistical and Machine Learning Methods for Time Series Forecasting in Data-Driven Logistics—A Simulation Study. *Entropy* **2025**, *27*, 25. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kurth, T.; Subramanian, S.; Harrington, P.; Pathak, J.; Mardani, M.; Hall, D.; Miele, A.; Kashinath, K.; Anandkumar, A. FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–11. [\[CrossRef\]](#)
- Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks. *Nature* **2023**, *619*, 533–538. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning Skillful Medium-Range Global Weather Forecasting. *Science* **2023**, *382*, 1416–1421. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, L.; Zhong, X.; Zhang, F.; Cheng, Y.; Xu, Y.; Qi, Y.; Li, H. FuXi: A Cascade Machine Learning Forecasting System for 15-Day Global Weather Forecast. *npj Clim. Atmos. Sci.* **2023**, *6*, 190. [\[CrossRef\]](#)
- Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T.R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. Probabilistic Weather Forecasting with Machine Learning. *Nature* **2025**, *637*, 84–90. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kochkov, D.; Yuval, J.; Langmore, I.; Norgaard, P.; Smith, J.; Mooers, G.; Klöwer, M.; Lottes, J.; Rasp, S.; Düben, P.; et al. Neural General Circulation Models for Weather and Climate. *Nature* **2024**, *632*, 1060–1066. [\[CrossRef\]](#) [\[PubMed\]](#)
- Rasp, S.; Hoyer, S.; Merose, A.; Langmore, I.; Battaglia, P.; Russell, T.; Sanchez-Gonzalez, A.; Yang, V.; Carver, R.; Agrawal, S.; et al. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *J. Adv. Model. Earth Syst.* **2024**, *16*, e2023MS004019. [\[CrossRef\]](#)
- Xu, B.; Yan, Z.; Li, Q.; Guo, Z.; Zheng, D.; Yao, W.; Wang, B.; Wang, Z.; Geng, Y. ASTAFN: Bridging the Gap Between Weather Foundation Models and Accurate Station-Level Forecasting. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2026; Volume 1, pp. 2494–2505. [\[CrossRef\]](#)
- Ben Bouallègue, Z.; Clare, M.C.A.; Magnusson, L.; Gascón, E.; Maier-Gerber, M.; Janoušek, M.; Rodwell, M.; Pinault, F.; Dramsch, J.S.; Lang, S.; et al. The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning-Based Weather Forecasts in an Operational-Like Context. *Bull. Am. Meteorol. Soc.* **2024**, *105*, E864–E883. [\[CrossRef\]](#)
- Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Zhang, C. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China, 10–16 August 2019; pp. 1907–1913. [\[CrossRef\]](#) [\[PubMed\]](#)

18. Jiang, R.; Wang, Z.; Yong, J.; Jeph, P.; Chen, Q.; Kobayashi, Y.; Song, X.; Fukushima, S.; Suzumura, T. Spatio-Temporal Meta-Graph Learning for Traffic Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2023; Volume 37, pp. 8078–8086. [[CrossRef](#)]
19. Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; Zhang, C. Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 753–763. [[CrossRef](#)]
20. Wang, Y.; Zhang, Z.; Pi, S.; Zhang, H.; Pi, J. Dual-Gated Graph Convolutional Recurrent Unit with Integrated Graph Learning (DG3L): A Novel Recurrent Network Architecture with Dynamic Graph Learning for Spatio-Temporal Predictions. *Entropy* **2025**, *27*, 99. [[CrossRef](#)] [[PubMed](#)]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
22. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2021; Volume 35, pp. 11106–11115. [[CrossRef](#)]
23. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. *arXiv* **2023**, arXiv:2211.14730.
24. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers Effective for Time Series Forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2023; Volume 37, pp. 11121–11128. [[CrossRef](#)]
25. Lim, B.; Arik, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [[CrossRef](#)]
26. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 22419–22430.
27. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 27268–27286. Available online: <https://proceedings.mlr.press/v162/zhou22g.html> (accessed on 3 September 2026).
28. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. *arXiv* **2023**, arXiv:2210.02186.
29. Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; Niu, Z. Frequency-Domain MLPs Are More Effective Learners in Time Series Forecasting. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 76656–76679. [[CrossRef](#)]
30. Ma, J.; Huang, Q.; Ma, H.; Wang, G.; Huang, S.; Zhou, Z.; Wang, P.; Wang, X.; Wang, B.; Wang, Y. PHAT: Modeling Period Heterogeneity for Multivariate Time Series Forecasting. In *Proceedings of the International Conference on Learning Representations*, Rio de Janeiro, Brazil, 23–27 April 2026. Available online: https://proceedings.iclr.cc/paper_files/paper/2026/hash/dbb5180957513805ebee787b8c66ac9-Abstract-Conference.html (accessed on 3 September 2026).
31. Lysenko, P.; Galyaev, A.; Berlin, L.; Babikov, V. Information Complexity of Time-Frequency Distributions of Signals in Detection and Classification Problems. *Entropy* **2025**, *27*, 998. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.