M^{2}WEATHER: A BENCHMARK FOR JOINT MULTI-STATION AND MULTI-VARIABLE WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Station weather forecasting is fundamentally shaped by both complex spatial dependencies across stations and strong physical coupling among weather variables. However, existing studies often consider these relationships separately and use different datasets and experimental settings, hindering systematic assessment of their individual and joint contributions. In this paper, we introduce M^{2}Weather, a benchmark for joint multi-station and multi-variable weather forecasting. Through multi-criteria quality control and station stratification, we collect 2,809 high-quality stations with 5 physically coupled weather variables across three spatial scales: France, Europe, and Global. This multi-scale design lets us examine whether conclusions persist from national to global station networks. We also introduce unified training and evaluation protocols to enable fair comparison of different station–variable modeling paradigms. To further examine the benefits of modeling station-variable relationships, we design a lightweight, plug-and-play adapter. With a trained weather forecasting model, this adapter can introduce missing station or variable relationships without retraining the model. This enables fair and efficient investigation of station–variable relationships. Systematic evaluation of 16 representative models shows the benefits of jointly modeling station and variable relationships. Completing missing relationships further reduces MSE for all adapted models on all three datasets. Together, these results identify the complementary information across stations and variables as an important resource for improving station weather forecasting.

1 INTRODUCTION

Station weather forecasting is a fundamental task in meteorology (Zhu et al., 2023), with important applications in agriculture, energy scheduling, transportation, and disaster early warning (Allen et al., 2025). However, accurate forecasting remains challenging because weather dynamics exhibit both complex spatial dependencies and multivariate coupling. On the one hand, geographical location, the propagation of atmospheric processes, and regional climate jointly give rise to substantial interactions across weather stations (Wu et al., 2023). On the other hand, weather variables, such as temperature, pressure, humidity, and wind speed, do not evolve independently; they continuously interact through atmospheric dynamics and thermodynamics (Ma et al., 2023). Therefore, station weather forecasting is naturally a prediction problem involving interactions among both *multiple stations* and *multiple variables*, where observations carry information about conditions at neighboring stations as well as related weather variables. Their usefulness may also vary with the spatial extent of the station network, motivating evaluation at multiple scales.

Despite the close interactions between stations and variables, existing research often treats these two dimensions separately. These methods can be classified into three categories: Single-Station Single-Variable (SSSV), Single-Station Multi-Variable (SSMV), and Multi-Station Single-Variable (MSSV) methods. SSSV methods (Fig. 1(a)) model individual stations and variables independently, as in channel-independent forecasting (Nie et al., 2023; Zeng et al., 2023). SSMV methods (Fig. 1(b)) capture dependencies between variables but typically lack explicit representations of station-level spatial structure (Liu et al., 2024a; Woo et al., 2024). MSSV methods (Fig. 1(c)) capture spatial relationships across stations while predicting different weather variables separately

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.

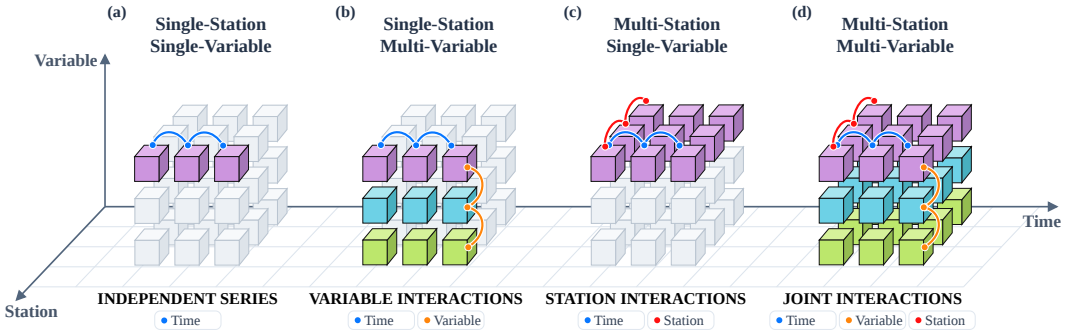

Figure 1: Four station–variable forecasting paradigms: (a) SSSV models each series independently; (b) SSMV captures interactions among variables within a station; (c) MSSV captures interactions across stations for one variable; and (d) MSMV models both. All four use temporal history, but differ in the station and variable context available to each forecast.

(Wu et al., 2023; Xu et al., 2025b). Methods jointly modeling Multi-Station Multi-Variable (MSMV, Fig. 1(d)) interactions remain relatively limited (Ma et al., 2023). More importantly, these methods are commonly evaluated using different datasets, interaction scopes, target variables, and experimental settings, hindering fair comparisons between modeling paradigms. We consequently lack a systematic evaluation for a fundamental question: **How do the individual and joint interactions among stations and variables affect station weather forecasting?**

Systematically investigating this question also places greater demands on data. First, reliable evaluation requires high-quality, complete station observations. High missing rates, poor data quality, or unstable station records can obscure underlying MSMV interactions, making model differences reflect data noise rather than forecasting capability. Second, weather variables have well-defined physical connections: their interactions are governed jointly by weather dynamics and thermodynamics. A benchmark for MSMV forecasting should therefore select weather quantities with clear physical coupling, so that variable interactions correspond to actual weather dynamics rather than statistical associations alone. Existing station datasets provide valuable foundations for forecasting research (Zhu et al., 2023; Han et al., 2024), but data and evaluation settings still lack a unified, systematic design that accounts for these two considerations.

To this end, we introduce **M^{2}Weather**, a **Multi-Station and Multi-Variable Benchmark** for station weather forecasting. This benchmark contains 2,809 high-quality stations with 5 physically coupled weather variables across three spatial scales: France, Europe, and Global. We further establish a unified, fair protocol for multi-station, multi-variable forecasting. To examine the effects of station and variable interactions, we develop a lightweight, plug-and-play adapter, which can introduce missing station or variable interactions to a trained forecasting model without retraining. Experiments of 16 baselines on M^{2}Weather show that MSMV methods achieve the best performance. With the adapter, we find that adding more missing station–variable interactions consistently improve performance for all adapted baselines. By adding all station–variable interactions, some non-MSMV methods can even achieve state-of-the-art performance. These findings support the value of MSMV modeling for station weather forecasting. Overall, our main contributions are summarized as follows:

- **A Benchmark for Joint Station–Variable Forecasting.** We construct a collection of high-quality station datasets with physically coupled weather variables and establish consistent training and evaluation protocols for investigating station–variable relationships in station weather forecasting.
- **Backpropagation-Free, Plug-and-Play Interaction Adapters.** We design lightweight station- and variable-interaction components that complement relationships missing from a baseline without retraining, enabling controlled analysis of each interaction dimension.
- **Systematic Evidence for MSMV Modeling.** Across the three datasets, the evaluated MSMV baseline achieves the best predictive performance, while every adapted backbone consistently improves with interaction completion. These findings highlight the value of jointly modeling station and variable relationships.

2 TAXONOMY OF STATION WEATHER FORECASTING

We classify methods by whether they explicitly model cross-station and cross-variable interactions. Shared parameters or pre-training alone do not imply either interaction.

2.1 SINGLE-STATION SINGLE-VARIABLE FORECASTING

Channel-independent methods learn temporal patterns from individual sequences, often sharing a predictor across channels. DLinear combines seasonal–trend decomposition with linear mappings (Zeng et al., 2023), while xPatch uses exponential decomposition and separate MLP/CNN streams to capture trend and seasonal components (Stitsyuk & Choi, 2025). PatchTST represents local temporal segments as patches, reducing attention costs and enabling longer look-back windows (Nie et al., 2023). This paradigm also extends to foundation models: Timer unifies heterogeneous data as single-series sequences for autoregressive next-token prediction (Liu et al., 2024b), and TimeMoE scales model capacity through sparse experts and multi-resolution forecasting heads (Shi et al., 2025). These methods provide temporal modeling baselines without explicitly coupling station or variable observations.

2.2 SINGLE-STATION MULTI-VARIABLE FORECASTING

Multivariate methods exploit dependencies among observed quantities within a station. iTransformer embeds each variable's history as a token and applies attention across variables (Liu et al., 2024a). TQNet combines periodically shifted learnable queries with sample-dependent keys and values to capture variable correlations (Lin et al., 2025). DUET jointly clusters temporal patterns and channels, using frequency-domain relationships and sparse attention to select relevant variables (Qiu et al., 2025). Among general forecasting models, Moirai supports varying numbers of variables through any-variate attention (Woo et al., 2024), while Timer-XL extends next-token prediction to multivariate sequences with TimeAttention that combines cross-variable dependencies and temporal causality (Liu et al., 2025). Applied within individual stations, these methods isolate the contribution of variable interactions without explicitly representing station-level spatial structure.

2.3 MULTI-STATION SINGLE-VARIABLE FORECASTING

This setting exploits relationships among irregularly distributed stations for a given weather variable. Corrformer combines temporal autocorrelation and spatial cross-correlation within a multi-scale tree to capture weather propagation (Wu et al., 2023). S^2Transformer uses intra-subgraph and inter-subgraph attention to capture local and regional relationships while supporting large station networks (Chen et al., 2026). CDPNet reconstructs a continuous spatial field from discrete observations and models its evolution through calibrated initialization and diffusive difference estimation (Xu et al., 2025b). Complementary approaches reduce computational costs: EasyST distills spatio-temporal graph models into an MLP with spatio-temporal prompts (Tang et al., 2024), while STELLA incorporates geographical coordinates and temporal embeddings into a lightweight MLP (Fu et al., 2025). Together, these methods span explicit spatial interaction and lightweight incorporation of spatio-temporal knowledge.

2.4 MULTI-STATION MULTI-VARIABLE FORECASTING

Joint modeling captures both within-station coupling and cross-station dependencies. HiSTGNN builds adaptive variable and station graphs with bidirectional exchange between levels (Ma et al., 2023). DIN combines cross-regional dynamics with higher-order variable interactions (Xu et al., 2025a), while StationPDE evolves a station-derived surface field using wind transport and coupled meteorological dynamics (Wang et al., 2026). These approaches illustrate distinct routes to explicit MSMV modeling, which remains comparatively underexplored.

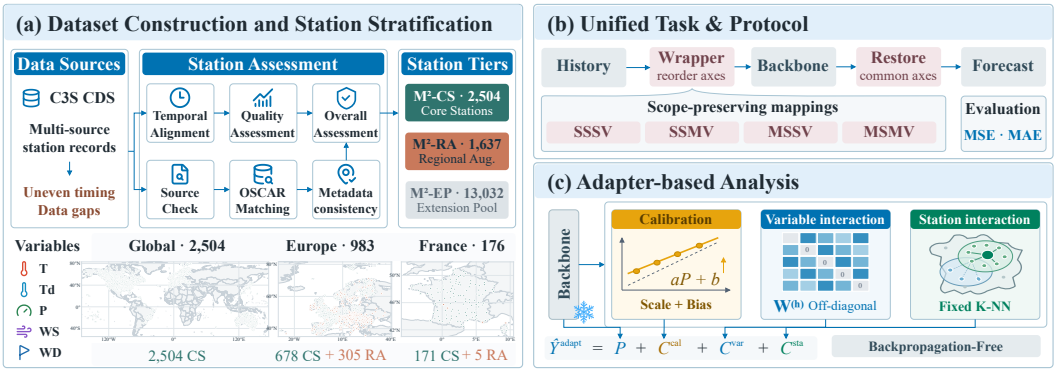

Figure 2: Overview of the M^{2}Weather benchmark and interaction analysis framework. (a) Dataset construction and station stratification based on temporal alignment, quality, provenance, and metadata. (b) Unified protocol for heterogeneous backbones, with model-specific wrappers preserving native interaction structures. (c) Adapter-based analysis on frozen benchmark backbones, with calibration and variable/station interaction corrections fitted by gradient-free ridge regression.

3 The M^{2}Weather Benchmark

3.1 Problem Definition

For a station network at a given spatial scale, let N denote the number of stations and V the number of weather variables. Let $x_{t,n,v}$ be the observation of variable v at station n and time t, and let L and H denote the history length and forecast horizon. The forecasting task is

$$\mathbf{X}_t = (x_{t-L+1:t,n,v})_{n,v}, \quad \mathbf{Y}_t = (x_{t+1:t+H,n,v})_{n,v}, \quad \widehat{\mathbf{Y}}_t = f_\theta(\mathbf{X}_t). \tag{1}$$

Here, $\mathbf{X}_t \in \mathbb{R}^{L\times N\times V}$ and $\mathbf{Y}_t \in \mathbb{R}^{H\times N\times V}$ are the historical input and future target tensors, respectively, while f_θ maps the former to the forecast $\widehat{\mathbf{Y}}_t$.

3.2 Dataset

3.2.1 Data Collection

M^{2}Weather is constructed from global land-surface observations provided by the Copernicus Climate Change Service through the Climate Data Store (CDS) (Copernicus Climate Change Service (C3S), 2026). The source product integrates station records from National Meteorological Services, aviation observing networks, and historical archives, together with quality flags, report types, and source identifiers. Despite its broad coverage, stations vary considerably in reporting frequency, variable availability, and temporal continuity, making the raw observations unsuitable for direct multi-station and multi-variable forecasting. We extract air temperature, dew-point temperature, station pressure, wind speed, and wind direction from 2017 to 2021. These variables describe the thermodynamic, moisture, pressure, and dynamical states of the near-surface atmosphere and exhibit meaningful physical coupling (Xu et al., 2024; Skamarock et al., 2019). The resulting candidate pool contains 17,173 stations and 43,824 hourly timestamps.

3.2.2 Station Assessment and Stratification

Reliable evaluation of station and variable interactions requires limiting the influence of missing observations, reporting interruptions, and inconsistent station records. We therefore assess stations along two complementary paths: temporal alignment followed by observation-quality assessment, and source checking followed by observing-network matching and metadata-consistency checks. Together, these provide the overall assessment used to organize stations into tiers, as illustrated in Fig. 2(a).

Temporal alignment. We first map irregular station reports to the nearest hourly timestamp in Coordinated Universal Time (UTC), retaining the temporally closest record when multiple reports

Dataset	Stations	Quality rule
Weather2K	2,130	$\leq 1\%$ missing; $\leq 1\%$ defaults/variable
GlobalTempWind	3,850	$\leq 20\%$ missing data
Weather5K	5,672	$> 90\%$ valid hourly records
M^{2}Weather	2,504/983/176	Multi-criteria station reliability tiers

Table 1: Reported station-screening rules. M^{2}Weather counts denote Global/Europe/France.

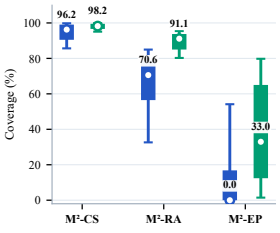

Figure 3: Pre-imputation coverage: minimum in blue and mean in green.

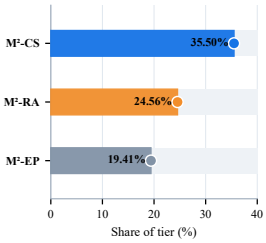

Figure 4: Verified current GBON affiliation.

are available for the same station, variable, and hour. For the released hourly series, we retain finite, quality-controlled observations within 15 minutes of the target hour, providing a standardized temporal basis for subsequent station assessment and dataset construction.

Quality assessment. Station suitability depends not only on overall data availability but also on how missing observations are distributed across variables and time. We jointly assess the long-term coverage of all five variables, monthly and seasonal stability, consecutive gap lengths, quality-control pass rates, and hourly-alignment rates. Detailed definitions and thresholds are provided in Appendix A.1.

Source and network checks. We use source identifiers, report types, and data-policy information in the source product to document the provenance of station records. We then match stations to the World Meteorological Organization (WMO) Observing Systems Capability Analysis and Review (OSCAR)/Surface catalogue (**?**). This match allows us to check the consistency of station identities, locations, and elevations, while providing information on operating status and observing-network affiliations. Together with historical observation quality, these source and network checks inform station selection and help interpret the resulting tiers. Matching rules and metadata definitions are provided in Appendix A.2.

Station tiers and dataset settings. The overall assessment organizes the 17,173 candidate stations into three mutually exclusive tiers. **M^{2}Weather Core Stations (M^2-CS)** contain 2,504 stations with high long-term completeness, stable monthly coverage, and consistent metadata. **M^{2}Weather Regional Augmentation Stations (M^2-RA)** contain 1,637 additional stations that satisfy the requirements for regional forecasting and improve local spatial coverage. The remaining 13,032 stations form the **M^{2}Weather Extension Pool (M^2-EP)** for future tasks and dataset extensions. Based on these tiers, we construct three dataset settings using the same variables, temporal range, and preprocessing protocol. **Global** contains all 2,504 core stations; **Europe** contains 678 core and 305 regional augmentation stations, totaling 983; and **France** contains 171 core and 5 regional augmentation stations, totaling 176. These settings provide different station configurations for evaluating station and variable interactions under a unified forecasting protocol.

3.2.3 Data Analysis

Comparison with existing station datasets. Table 1 compares M^{2}Weather with representative hourly station-weather datasets (Zhu et al., 2023; Wu et al., 2023; Han et al., 2024). The reported screening rules for these datasets primarily rely on one or two sequence-based availability thresholds. M^{2}Weather additionally combines multivariate and temporal quality checks with verified station identities and observing-network affiliations to define reliability tiers.

Quality and network evidence. Fig. 3 summarizes station coverage before imputation. The median minimum-variable and mean-variable coverage values are 96.24% and 98.23% for M^2-CS, compared with 70.58% and 91.11% for M^2-RA. Fig. 4 provides complementary observing-network evidence. The proportion of stations with both a verified identity in the WMO OSCAR/Surface catalogue and current Global Basic Observing Network (GBON) affiliation decreases from 35.50% in M^2-CS to 24.56% in M^2-RA and 19.41% in M^2-EP. Together, these results show that the proposed tiers capture both historical multivariate usability and externally documented observing-network roles.

3.3 Unified Training and Evaluation Protocols

Data splits and standardization. For each spatial scale, we split the data chronologically into training, validation, and test sets with a ratio of 60%/10%/30%. Both historical and forecast windows remain entirely within their assigned split. For each variable, we compute the mean and standard deviation over all stations in the training set of the corresponding spatial scale and use these fixed statistics for validation and testing.

Model wrappers. As illustrated in Fig. 2(b), we provide unified wrappers that rearrange the station and variable dimensions to match each backbone's native input format and restore predictions to the common output layout. For SSMV models, stations are folded into the batch dimension, whereas for MSSV models, variables are folded into the batch dimension. Parameters are shared across the resulting samples, allowing each trainable baseline to be trained once per dataset and random seed, without training a separate model for each variable or station. This design supports consistent comparison while preserving each model's intended interaction scope.

Training and hyperparameters. We select four variables from the M^{2}Weather datasets for the forecasting task: air temperature, relative humidity, station pressure, and wind speed. All baselines use 48 hours of observations of these variables to predict their values over the next 72 hours. We run all experiments on NVIDIA A100 GPUs (40 GB), with three runs for supervised models using seeds 2024, 2025, and 2026. All trained or fine-tuned backbones use Adam, MSE as the forecasting loss, a common initial learning rate of 10^{-4}, and the type1 learning-rate decay schedule. Model-specific auxiliary losses and regularization are retained where applicable. Timer, TimeMoE, and Moirai use zero-shot inference. Batch sizes depend on the backbone and spatial scale and are adjusted to available GPU memory by the execution scheduler. Training lasts at most 30 epochs, and checkpoints are selected using each model's validation objective. Early-stopping patience is 10 for France and 5 for Europe and Global. Model-specific hyperparameters, sampling settings, and pretrained inference configurations are detailed in Appendix B.1.

Evaluation metrics. We report mean squared error (MSE) and mean absolute error (MAE). Let Ω contain all evaluated forecast-origin, horizon, station, and variable tuples, and let $e_i = \widehat{Y}_i - Y_i$ denote the prediction error for tuple i. The metrics are

$$\mathrm{MSE} = \frac{1}{|\Omega|}\sum_{i\in\Omega} e_i^2, \qquad \mathrm{MAE} = \frac{1}{|\Omega|}\sum_{i\in\Omega} |e_i|. \tag{2}$$

Overall metrics are computed in standardized space with equal weight across variables; per-variable MSE and MAE in the same standardized space are reported in Appendix C.3. All methods use the same test origins, forecast steps, and stations, with predictions from overlapping windows evaluated separately.

4 Plug-and-Play Adapters for Station-Variable Interaction Analysis

To investigate how joint station–variable modeling benefits weather forecasting, we design lightweight, backpropagation-free post-processing adapters that complement interactions missing from existing backbones. The adapters introduce variable and station information directly into the forecast outputs, enabling a backbone to access both interaction dimensions without modifying its architecture or retraining its parameters. Their coefficients admit a closed-form solution, making interaction completion efficient across the diverse model families in M^{2}Weather. The adapter-based interaction analysis is summarized in Fig. 2(c).

Output-level residual correction. Let $\mathbf{P}_t = f_\theta(\mathbf{X}_t) \in \mathbb{R}^{H\times N\times V}$ denote the frozen backbone forecast at origin t. The adapter operates in the dataset-standardized space and adds calibration, variable-interaction, and station-interaction corrections:

$$\widehat{\mathbf{Y}}_t^{\mathrm{adapt}} = \mathbf{P}_t + \mathbf{C}_t^{\mathrm{cal}} + \mathbf{C}_t^{\mathrm{var}} + \mathbf{C}_t^{\mathrm{sta}}. \tag{3}$$

The backbone remains fixed in evaluation mode, and the enabled interaction branches process the same original predictions in parallel. We use $C^{\mathrm{cal}}_{t,h,n,v} = a_{h,v}P_{t,h,n,v} + b_{h,v}$, where the learnable scale and bias corrections are shared across stations. This term accounts for each variable's own

amplitude and bias errors, allowing the interaction branches to focus on cross-variable and cross-station information rather than having to absorb these basic calibration effects.

Variable-interaction adapter. For each station n and forecast step h, we normalize the prediction vector across its V variables and denote the result by $\widetilde{\mathbf{P}}_{t,h,n,:}$. We then aggregate the normalized predictions of other variables through a learnable off-diagonal mapping:

$$C^{\mathrm{var}}_{t,h,n,v} = \sum_{u \neq v} W^{(h)}_{u,v} \widetilde{P}_{t,h,n,u}, \qquad W^{(h)}_{v,v} = 0. \tag{4}$$

Here, $W^{(h)}_{u,v}$ controls how variable u contributes to the correction of variable v at forecast step h. The mapping is shared across stations and introduces variable interactions within each station, without exchanging information between stations.

Station-interaction adapter. We construct a fixed set $\mathcal{N}(n)$ of eight nearest neighbors for each station, excluding itself, using great-circle angular distances d_{nj} computed from geographical coordinates. Let τ_n be the median neighbor distance with a positive lower bound. The row-normalized adjacency matrix is

$$A_{nj} = \frac{\exp(-d_{nj}/\tau_n)}{\sum_{k \in \mathcal{N}(n)} \exp(-d_{nk}/\tau_n)}, \qquad j \in \mathcal{N}(n), \tag{5}$$

with all other entries zero. For each forecast step and variable, we normalize predictions across stations and denote the resulting vector by $\mathbf{Z}_{t,h,:,v}$. The station correction combines one- and two-hop spatial contrasts:

$$\mathbf{C}^{\mathrm{sta}}_{t,h,:,v} = \sum_{r=1}^{2} \gamma^{(r)}_{h,v} \big(\mathbf{A}^r \mathbf{Z}_{t,h,:,v} - \mathbf{Z}_{t,h,:,v} \big). \tag{6}$$

Each $\gamma^{(r)}_{h,v}$ is a learnable coefficient shared across stations. Its magnitude and sign determine the strength and direction of the correction associated with the r-hop neighborhood, adapting the influence of local and broader spatial patterns to each variable and forecast lead time. The contrast measures how a station's prediction differs from its surrounding field. This branch exchanges station information separately for each variable, and neither interaction branch mixes forecast steps.

Closed-form solution. With backbone predictions fixed, the adapter correction is linear in all learnable coefficients. Under squared error, fitting therefore reduces to a quadratic problem that can be solved directly by ridge regression. For each forecast-step–variable pair (h, v), we stack $M = |\mathcal{T}_{\mathrm{train}}|N$ training origin–station samples into a feature matrix $\mathbf{\Phi}_{h,v}$ and construct a residual vector $\mathbf{r}_{h,v}$ with entries $Y_{t,h,n,v} - P_{t,h,n,v}$, where $\mathcal{T}_{\mathrm{train}}$ denotes the training forecast origins. Let $\boldsymbol{\beta}_{h,v}$ collect $a_{h,v}$, $b_{h,v}$, and the enabled interaction coefficients. We solve

$$\boldsymbol{\beta}^{\star}_{h,v} = \arg\min_{\boldsymbol{\beta}} \ \frac{1}{M} \|\mathbf{\Phi}_{h,v}\boldsymbol{\beta} - \mathbf{r}_{h,v}\|_2^2 + \lambda \boldsymbol{\beta}^{\top} \mathbf{D} \boldsymbol{\beta}, \tag{7}$$

whose normal equations yield the following expression when the coefficient matrix is nonsingular:

$$\boldsymbol{\beta}^{\star}_{h,v} = \big(\mathbf{\Phi}^{\top}_{h,v} \mathbf{\Phi}_{h,v} + M\lambda \mathbf{D} \big)^{-1} \mathbf{\Phi}^{\top}_{h,v} \mathbf{r}_{h,v}. \tag{8}$$

Here, $\mathbf{D}$ leaves the calibration intercept unpenalized, and $\lambda \geq 0$ is selected by overall standardized validation MSE. We solve the normal equations numerically without explicitly forming the displayed inverse; the expression does not assume that every candidate produces an invertible matrix. We accumulate feature sufficient statistics batch by batch and perform a single small regression solve per (h, v) for each candidate λ, reusing the same statistics across candidates. This yields a lightweight fitting procedure without iterative gradient descent or backpropagation. Appendix B.2 details the feature construction, streaming solution, and parameter-efficiency analysis. In practice, branches are fitted according to the interactions to be completed. The full-completion configurations add station interactions to SSMV, variable interactions to MSSV, and both to SSSV, supplying both information dimensions at the output level.

5 Observations and Findings

We evaluate 16 representative backbones spanning the SSSV, SSMV, MSSV, and MSMV paradigms on France, Europe, and Global (Table 2). They cover linear models, Transformers, graph neural

Table 2: Baseline coverage under the four station–variable modeling paradigms in M^{2}Weather.

Paradigm	Baselines
SSSV	DLinear (Zeng et al., 2023), xPatch (Stitsyuk & Choi, 2025), PatchTST (Nie et al., 2023), Timer (Liu et al., 2024b), TimeMoE (Shi et al., 2025)
SSMV	iTransformer (Liu et al., 2024a), TQNet (Lin et al., 2025), DUET (Qiu et al., 2025), Moirai (Woo et al., 2024), Timer-XL (Liu et al., 2025)
MSSV	Corrformer (Wu et al., 2023), S^{2}Transformer (Chen et al., 2026), CDPNet (Xu et al., 2025b), EasyST (Tang et al., 2024), STELLA (Fu et al., 2025)
MSMV	HiSTGNN (Ma et al., 2023)

Table 3: MSE ($48 \rightarrow 72$ hours). Δ is the relative change after adaptation; negative values indicate lower error. HiSTGNN is reported without an adapter. Underlining marks the best native backbone; bold marks the best displayed result (ties at four decimals). Per-variable MSE, MAE, and standard deviations are in Appendix C.3.

		France			**Europe**			**Global**		
Paradigm	Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
SSSV	DLinear	0.4977	0.4605	−7.49	0.4481	0.4087	−8.79	0.4508	0.4172	−7.45
	xPatch	0.3974	0.3642	−8.36	0.3509	0.3280	−6.53	0.3587	0.3371	−6.04
	PatchTST	0.3905	0.3601	−7.80	0.3472	0.3266	−5.92	0.3540	0.3341	−5.61
	Timer	0.5211	0.4445	−14.70	0.4392	0.3845	−12.47	0.4547	0.3986	−12.33
	TimeMoE	0.4354	0.3845	−11.69	0.3798	0.3435	−9.56	0.3943	0.3559	−9.73
SSMV	iTransformer	0.3502	0.3335	−4.79	0.3256	0.3120	−4.18	0.3233	0.3119	−3.51
	TQNet	0.3704	0.3484	−5.93	0.3472	0.3274	−5.70	0.3506	0.3333	−4.94
	DUET	0.4052	0.3737	−7.77	0.3905	0.3646	−6.63	0.3640	0.3420	−6.06
	Moirai	0.4717	0.3817	−19.08	0.4124	0.3438	−16.63	0.4326	0.3585	−17.13
	Timer-XL	0.3483	0.3331	−4.36	0.3132	0.3031	−3.22	0.3177	0.3079	−3.10
MSSV	Corrformer	0.3684	0.3417	−7.24	0.3099	0.2988	−3.60	0.3189	0.3079	−3.44
	S^{2}Transformer	0.3776	0.3650	−3.33	0.3332	0.3252	−2.39	0.3158	0.3148	−0.32
	CDPNet	0.3311	0.3248	−1.91	0.2940	**0.2877**	−2.15	0.2985	**0.2961**	−0.81
	EasyST	0.3442	0.3398	−1.27	0.3108	0.3086	−0.71	0.3101	0.3087	−0.45
	STELLA	0.3878	0.3690	−4.84	0.3732	0.3520	−5.68	0.4285	0.3931	−8.26
MSMV	HiSTGNN	**0.3079**	–	–	0.2908	–	–	**0.2961**	–	–

networks, and pretrained time-series foundation models, enabling comparisons across interaction scopes and model families. For MSMV, we evaluate HiSTGNN using its publicly available implementation. The implementations of DIN and StationPDE are not publicly released, so they cannot be included in the reproducible comparison under our unified protocol. Using 48 hours of history, each model predicts the next 72 hours. The main text reports mean standardized MSE, averaged equally over the four target variables; complete per-variable MSE, MAE, and standard deviations are provided in Appendix C.3.

5.1 Does Joint Station–Variable Modeling Improve Forecasting?

We first compare the native backbone results in Table 3 and Figure 5. Across all three datasets, SSSV has the highest group-mean MSE, whereas the native MSMV reference, HiSTGNN, achieves the lowest MSE among all 16 backbones. On France, for example, the SSSV mean is 0.4484, compared with 0.3892 for SSMV, 0.3618 for MSSV, and 0.3079 for HiSTGNN. The same ordering holds on Europe and Global. This consistent pattern supports the value of jointly exploiting station and variable relationships over treating each series independently. Comparing the two intermediate paradigms reveals a further pattern: MSSV outperforms SSMV in group-mean MSE at every spatial scale. Their respective means are 0.3618 versus 0.3892 on France, 0.3242 versus 0.3578 on Europe, and 0.3344 versus 0.3576 on Global. These results suggest that information exchanged across stations may be particularly valuable for station weather forecasting. HiSTGNN is the only dedicated MSMV model among the evaluated baselines. Its performance, together with the interaction-completion results below, motivates broader investigation of architectures that explicitly model both station and variable relationships.

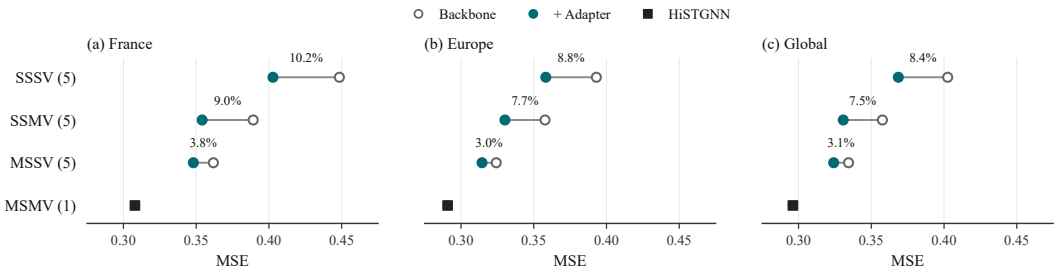

Figure 5: MSE by modeling paradigm, averaged over models. Connected markers compare native and adapted predictions; percentages denote reductions in group means. Parentheses indicate group sizes. HiSTGNN is the native MSMV reference, without an adapter.

5.2 Does Completing Missing Interactions Improve Forecasting?

Table 3 shows that the adapter improves all 15 SSSV, SSMV, and MSSV backbones on every dataset: all 45 model–dataset pairs achieve lower MSE, with relative reductions of 0.32–19.08%. For the 12 stochastic backbones with three independent runs, seed-level paired tests remain significant after within-dataset Holm correction in 34 of 36 comparisons (Appendix C.2). Thus, interaction completion benefits diverse backbones while keeping their parameters fixed, providing complementary evidence for jointly modeling the station and variable dimensions. The group-level results in Figure 5 further reveal where these benefits are largest. SSSV achieves the greatest relative decrease in group-mean MSE at each scale: 10.2%, 8.8%, and 8.4% on France, Europe, and Global. SSMV follows with reductions of 9.0%, 7.7%, and 7.5%, exceeding the corresponding MSSV reductions of 3.8%, 3.0%, and 3.1%. The larger gains for the station-independent groups are consistent with the native comparisons above, suggesting substantial value in adding spatial information. Appendix C.4 expands the group summaries into individual backbone comparisons.

To examine whether these gains persist without calibration, we disable the calibration branch and refit only the enabled interaction branches on France and Global. Table 22 in Appendix C.5 shows that this interaction-only variant improves all 30 evaluated model–dataset pairs. For example, Moirai's France MSE decreases from 0.4717 to 0.4094 without calibration, and to 0.3817 with the full adapter. Improvements therefore persist when the correction contains only interaction terms, while calibration provides additional benefits. Together, the native comparisons and frozen-backbone results support exploiting both station and variable information for accurate station forecasts. Prediction examples on Europe (Appendix C.10, Figures 9–12) provide a complementary view: adapted forecasts better track sustained temperature increases, humidity decreases, and rising wind speeds, while most pressure forecasts align more closely with observations. These examples illustrate improvements in forecast levels and trends across backbones.

6 Conclusion

We introduced M^{2}Weather to systematically assess the individual and joint roles of station and variable interactions in weather forecasting. The benchmark combines quality-controlled datasets at three spatial scales with unified protocols for comparing four modeling paradigms. Our lightweight, plug-and-play adapters complement missing interactions in frozen backbone forecasts through closed-form ridge regression, requiring no backpropagation. Evaluations of 16 representative models support joint station–variable modeling and highlight the value of spatial information. All 15 adapted backbones achieve lower MSE on all three datasets, and interaction-only ablations show that gains persist without calibration. Together, these findings call for greater attention to forecasting architectures that explicitly capture both station and variable interactions. M^{2}Weather provides a foundation for extending this analysis to richer joint architectures, additional forecast horizons, and extreme-weather conditions.

AI USE STATEMENT

Generative AI assisted the authors with preliminary exploration, literature review, and manuscript polishing and editing. The authors remain fully responsible for the accuracy of the text, the cited references, and the reported research.

REFERENCES

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. End-to-end data-driven weather prediction. *Nature*, 641:1172–1179, 2025. doi: 10.1038/s41586-025-08897-0. URL https://www.nature.com/articles/s41586-025-08897-0.

Hongyi Chen, Xiucheng Li, Xinyang Chen, Yun Cheng, Jing Li, Kehai Chen, and Liqiang Nie. S^2Transformer: Scalable structured transformers for global station weather forecasting. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=AL2VnKno5n.

Copernicus Climate Change Service (C3S). Global land surface atmospheric variables from 1718 to present from comprehensive in-situ observations. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2026. URL https://doi.org/10.24381/cds.cf5f3bac. Accessed: 16 September 2026.

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 *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/7691484a7a35d5e2742279c1d926b778-Paper-Conference.pdf.

Tao Han, Song Guo, Zhenghao Chen, Wanghan Xu, and Lei Bai. Weather-5k: A large-scale global station weather dataset towards comprehensive time-series forecasting benchmark. *arXiv preprint arXiv:2406.14399*, 2024. URL https://arxiv.org/abs/2406.14399v1.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Proceedings of the International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2505.12917.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024a. URL https://arxiv.org/abs/2310.06625.

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*, 2024b. URL https://arxiv.org/abs/2402.02368.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KMCJXjlDDr.

Minbo Ma, Peng Xie, Fei Teng, Bin Wang, Shenggong Ji, Junbo Zhang, and Tianrui Li. HiSTGNN: Hierarchical spatio-temporal graph neural network for weather forecasting. *Information Sciences*, 648:119580, 2023. doi: 10.1016/j.ins.2023.119580. URL https://doi.org/10.1016/j.ins.2023.119580.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1185–1196, 2025. URL https://arxiv.org/abs/2412.10859.

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 *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.16040.

W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, Z. Liu, J. Berner, W. Wang, J. G. Powers, M. G. Duda, D. M. Barker, and X.-Y. Huang. A description of the advanced research WRF version 4. NCAR Technical Note NCAR/TN-556+STR, National Center for Atmospheric Research, 2019. URL https://doi.org/10.5065/1dfh-6p97.

Artyom Stitsyuk and Jaesik Choi. xPatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2412.17323.

Jiabin Tang, Wei Wei, Lianghao Xia, and Chao Huang. EasyST: A simple framework for spatio-temporal prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2220–2229. ACM, 2024. doi: 10.1145/3627673.3679749. URL https://doi.org/10.1145/3627673.3679749.

Xiao Wang, Changjian Chen, Rongwen Li, Hongwu Liu, Kun Fang, and Zhuo Tang. StationPDE: Station-oriented surface pde learning for multi-station multivariate weather forecasting, 2026. URL https://arxiv.org/abs/2609.22123.

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*, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Haixu Wu, Hang Zhou, Mingsheng Long, and Jianmin Wang. Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, 5:602–611, 2023. doi: 10.1038/s42256-023-00667-9. URL https://www.nature.com/articles/s42256-023-00667-9.

Chujie Xu, Yinkai Liu, Yajun Gao, Xiaotong Zhu, Yudie Wang, Yong Han, and Yan Wu. DIN: Dynamical interaction network for multi-station multi-variable weather prediction. In *Generalizing from Limited Resources in the Open World: Third International Workshop, GLOW 2025, Held in Conjunction with IJCAI 2025, Proceedings*, volume 2640 of *Communications in Computer and Information Science*, pp. 95–107. Springer, 2025a. doi: 10.1007/978-981-95-0988-1_7. URL https://doi.org/10.1007/978-981-95-0988-1_7.

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, 2025b. doi: 10.24963/ijcai.2025/747. URL https://doi.org/10.24963/ijcai.2025/747.

Wanghan Xu, Fenghua Ling, Wenlong Zhang, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. Generalizing weather forecast to fine-grained temporal scales via physics-ai hybrid modeling. In *Proceedings of the Advances in Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/abs/2205.13504.

Xun Zhu, Yutong Xiong, Ming Wu, Gaozhen Nie, Bin Zhang, and Ziheng Yang. Weather2K: A multivariate spatio-temporal benchmark dataset for meteorological forecasting based on real-time observation data from ground weather stations. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2704–2722, 2023. URL https://proceedings.mlr.press/v206/zhu23a.html.

APPENDIX OVERVIEW

This appendix provides supplementary details on dataset construction, benchmark implementation, and experimental results. Section A covers quality criteria and station tiering, station identity and provenance with network matching, imputation masks and target-variable construction, and spatial subset definitions with additional statistics. Section B presents model hyperparameters across datasets, adapter fitting, and parameter efficiency. Section C presents comprehensive forecasting results, ablation and sensitivity analyses, computational efficiency evaluations, and qualitative prediction visualizations.

A DATASET CONSTRUCTION DETAILS

A.1 QUALITY CRITERIA AND STATION-TIERING THRESHOLDS

Physical rationale for variable selection. We select air temperature (T), dew-point temperature (T_d), station pressure (p), wind speed (s), and wind-from direction (ϕ) to describe complementary aspects of the near-surface thermal, moisture, pressure, and flow states. These aspects are physically connected: pressure gradients contribute to wind evolution, winds transport heat and moisture between locations, and changes in water vapour can be accompanied by latent heating. WRF represents these processes through coupled momentum, thermodynamic, and moisture equations (Skamarock et al., 2019), while WeatherGFT uses their coupled evolution to guide physics-based forecasting (Xu et al., 2024). The following simplified atmospheric balances illustrate these interactions:

$$\frac{D\mathbf{u}_h}{Dt} + f\hat{\mathbf{k}} \times \mathbf{u}_h = -\frac{1}{\rho}\nabla_h p + \mathbf{F}_h, \tag{9}$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{10}$$

$$c_p \frac{DT}{Dt} - \frac{1}{\rho}\frac{Dp}{Dt} = Q_{\mathrm{diab}}, \tag{11}$$

$$\frac{Dq}{Dt} = S_q, \qquad p = \rho R_d T_v, \quad T_v \simeq T(1 + 0.61q). \tag{12}$$

Here $\mathbf{u}_h = (u, v)$, $\mathbf{u} = (u, v, w)$, and $D/Dt = \partial_t + \mathbf{u} \cdot \nabla$. The material derivatives capture wind-driven transport of temperature and moisture; the pressure-gradient term links pressure to the flow; and moisture changes associated with phase transitions also affect the heating term Q_{diab}.

The selected station observations provide corresponding near-surface quantities:

$$\begin{aligned} (u, v) &= (-s\sin\phi,\ -s\cos\phi), & e &= e_s(T_d), \\ \mathrm{RH} &= 100\frac{e_s(T_d)}{e_s(T)}, & q &= \frac{\varepsilon e}{p - (1-\varepsilon)e}, \end{aligned} \tag{13}$$

where e_s is saturation vapour pressure, relative humidity is expressed as a percentage, and $\varepsilon = R_d/R_v$. Wind speed and direction jointly describe the horizontal flow, whereas temperature, dew point, and pressure jointly characterize near-surface moisture. Although station observations alone do not resolve the full three-dimensional state required to integrate the atmospheric equations, they capture connected aspects of the same physical system. These local thermodynamic relationships and wind-mediated links between locations motivate evaluating variable and station interactions together.

Hourlyization and counting convention. Reports are assigned to their nearest UTC hour, with exact half-hour timestamps rounded forward. Reports mapped outside their source calendar month are discarded, and duplicate reports for the same station, variable, and hour are resolved by retaining the report closest to the target hour. Coverage and gap statistics are computed from these de-duplicated hourly cells before imputation. Quality-control (QC) pass rates and hourly-alignment rates are evaluated separately, preventing observation availability, quality flags, and reporting regularity from being conflated.

Quality metrics. The evaluation period contains 43,824 hours and 60 calendar months. For each of the five variables, five-year coverage is defined as the number of occupied hourly cells divided by 43,824. Monthly coverage statistics include the lower quantiles of the minimum, mean, and fourth-highest coverage across variables. Temporal continuity is measured using upper quantiles of monthly maximum gap lengths, including gaps at month boundaries. We additionally compute weighted QC pass rates, the proportion of hourly cells represented by reports within 15 minutes of the target hour, and the longest station-wide outage, where an hour is considered available if any target variable is observed.

A month is metadata-stable when its longitude and latitude spans do not exceed 0.01° and its valid elevation span does not exceed 10 m. A station is marked as having an identity conflict when its five-year coordinate or elevation variation exceeds these limits, or when no valid location is available. Each variable is also required to occur in all four meteorological seasons. All monthly quantiles use linear interpolation over the 60 calendar months.

Stations satisfying all M^2-CS criteria form the core tier. After excluding these stations, those satisfying all M^2-RA criteria form the regional augmentation tier. The remaining stations are retained in the M^2-EP extension pool. This procedure yields 2,504 M^2-CS, 1,637 M^2-RA, and 13,032 M^2-EP stations.

A.2 Station Identity, Provenance, and Network Matching

Source provenance. Station identities are first linked to the C3S station inventory through their primary station identifiers. All 17,173 candidate stations are represented in the inventory. For the observations retained by the strict hourly processing pipeline, each source identifier is resolved against the official C3S source table. We preserve the associated acknowledgement text, report-type composition, and data-policy information as complementary provenance metadata.

OSCAR/Surface matching. We cross-reference the station catalogue with the World Meteorological Organization (WMO) Observing Systems Capability Analysis and Review (OSCAR)/Surface catalogue (**?**). Matching considers all available WMO Integrated Global Observing System (WIGOS) station identifiers and uses three evidence levels. A match is accepted when: (1) the complete WIGOS identifier matches exactly; (2) the local identifier agrees, the stations are within 10 km, and country, station name, or elevation provides additional support; or (3) stations belong to the same country and have either an exact name within 5 km, a name-similarity score of at least 0.90 within 1 km, or a separation below 0.25 km with an elevation difference below 100 m. Coordinate-only matches within 5 km are retained as possible candidates but are not used to assign official network attributes.

For accepted matches, we record station identity, location, operating status, and membership in the Global Basic Observing Network (GBON), Regional Basic Observing Network (RBON), and Global Climate Observing System (GCOS) Surface Network (GSN).

Network memberships are taken from the OSCAR/Surface snapshot retrieved on 9 September 2026 and therefore provide current external station context. Historical observation quality, C3S source provenance, and WMO observing-network information together constitute the evidence used to interpret the station tiers.

A.3 Imputation Masks and Target-Variable Construction

Short-gap interpolation. A missing interval is treated as a short internal gap only when it is bounded by valid observations from the same station and variable and contains no more than 12 consecutive hours. Air temperature, dew-point temperature, station pressure, and wind speed are linearly interpolated between the two endpoints. For wind direction, we interpolate along the shortest circular arc. Given endpoint directions d_l and d_r, the angular displacement is

$$\Delta_d = \big((d_r - d_l + 180^\circ) \bmod 360^\circ\big) - 180^\circ, \tag{14}$$

and an intermediate value at relative position $\alpha \in (0, 1)$ is computed as

Table 4: Station-tiering criteria. A dash indicates that the criterion is not used for the corresponding tier.

Criterion	M^2-CS	M^2-RA
Five-year coverage of every variable	≥ 0.82	≥ 0.20
Five-year mean coverage	–	≥ 0.65
Monthly minimum coverage, P_{10}	≥ 0.60	–
Monthly fourth-highest coverage, P_{20}	–	≥ 0.50
Monthly mean coverage, P_{20}	–	≥ 0.50
Monthly maximum five-variable gap, P_{90}	≤ 168 h	–
Monthly maximum existing-variable gap, P_{80}	–	≤ 72 h
Five-year weighted QC pass rate	≥ 0.98	≥ 0.95
Monthly QC quantile	$P_{10} \geq 0.95$	$P_{20} \geq 0.95$
Five-year hourly-alignment rate	≥ 0.90	≥ 0.75
Monthly alignment quantile	$P_{10} \geq 0.75$	$P_{20} \geq 0.75$
Metadata-stable month ratio	≥ 0.90	≥ 0.80
Maximum station-wide outage	≤ 2160 h	≤ 2880 h
All variables present in all seasons	Required	Required
Identity conflict	Not allowed	Not allowed

$$d(\alpha) = (d_l + \alpha\Delta_d) \bmod 360^\circ. \tag{15}$$

This formulation correctly interpolates across the $0^\circ/360^\circ$ boundary.

ERA5-based completion. After short-gap interpolation, all remaining missing values, including internal gaps longer than 12 hours and gaps at sequence boundaries, are matched at the same UTC hour to the nearest 0.25° grid cell in ERA5, the fifth-generation atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts. Using reanalysis to complete station records follows established dataset-construction practice: WEATHER-5K also applies short-gap interpolation before using ERA5 to fill remaining gaps (Han et al., 2024).

$$T = \mathrm{t2m} - 273.15, \qquad T_d = \mathrm{d2m} - 273.15, \tag{16}$$

$$P = \mathrm{sp}/100, \qquad WS = \sqrt{\mathrm{u10}^2 + \mathrm{v10}^2}, \tag{17}$$

$$WD = \left[\mathrm{atan2}(-\mathrm{u10}, -\mathrm{v10})\frac{180^\circ}{\pi}\right] \bmod 360^\circ, \tag{18}$$

where T and T_d are measured in degrees Celsius, P in hectopascals, WS in meters per second, and WD is the meteorological wind-from direction.

As safeguards, the implementation can fall back sequentially to a nearby hour at the same ERA5 grid cell, a neighboring grid cell at the same hour, an inverse-distance-weighted estimate from neighboring stations, the nearest valid observation from the same station, and finally the cross-station median. None of these fallback stages were activated in the released datasets: all remaining gaps were filled by the nearest ERA5 grid cell at the corresponding hour.

Value-source masks. For every released value, we store a source code in `value_source.npy`. The mask has shape $[N, 5, T]$, corresponding to station, meteorological variable, and time. Its dynamic-variable order is air temperature, dew-point temperature, wind speed, station pressure, and wind direction.

Released variables and arrays. Each dataset is stored as a `float32` array with shape $[N, 8, T]$. The first three channels are latitude, longitude, and elevation, repeated along the temporal dimension. The remaining channels are air temperature, dew-point temperature, wind speed, station pressure, and wind direction. Wind direction is retained in its original meteorological representation over $[0^\circ, 360^\circ)$ rather than being replaced by wind-vector components.

Table 5: Per-value source codes stored in the imputation mask.

Code	Source	Present in release
0	Unresolved missing value	No
1	Quality-controlled exact-hour observation	Yes
2	Quality-controlled observation within 15 minutes	Yes
3	Same-station interpolation of a gap ≤ 12 h	Yes
4	Nearest ERA5 grid cell at the same hour	Yes
5	Same ERA5 grid cell at a nearby hour	No
6	Neighboring ERA5 grid cell at the same hour	No
7	Neighboring-station estimate	No
8	Nearest valid time from the same station	No
9	Cross-station median	No

Table 6: Percentage of values completed by either station interpolation or ERA5.

Dataset	T	T_d	WS	P	WD	Overall
Global	1.392	1.953	1.578	2.343	5.576	2.568
Europe	2.608	3.216	3.630	11.242	7.499	5.639
France	0.585	0.650	0.735	1.266	1.381	0.923

Four-variable forecasting task. After short-gap interpolation and completion of any remaining gaps, we derive relative humidity from the processed air and dew-point temperatures. Before standardization, we compute $z = 17.625T_d/(243.04 + T_d) - 17.625T/(243.04 + T)$ with both temperatures in degrees Celsius, clip z to $[-50, 50]$, and set $\mathrm{RH} = \mathrm{clip}(100\exp(z), 0, 100)$. Thus, the released five-variable observations retain dew-point temperature, while the forecasting task uses its derived relative humidity together with air temperature, station pressure, and wind speed. Wind speed and direction describe the magnitude and orientation of the same horizontal wind field. We retain wind direction in the released data but select wind speed as the wind-related target in the present four-variable task; joint prediction of both wind quantities is left for future extensions.

All released arrays contain finite values only. The source masks are retained to support observed-only evaluation, imputation-aware training, and sensitivity analyses of the completion strategy.

Dataset standardization. For each spatial scale and target variable v, we compute $\mu_v = C_v^{-1}\sum_i x_{i,v}$ and $\sigma_v = \sqrt{\max(C_v^{-1}\sum_i x_{i,v}^2 - \mu_v^2, 10^{-12})}$ over all training station–hour values, with equal weight per value and population variance. The maximum ensures $\sigma_v \geq 10^{-6}$. Inputs and targets are transformed as $(x_{i,v} - \mu_v)/\sigma_v$, with these statistics frozen for validation and testing. The completed training arrays, including station-interpolated and ERA5-filled values, contribute to these statistics; relative humidity is derived beforehand.

Temporal boundaries. All three spatial scales share the same 43,824-hour UTC timeline and the boundaries in Table 7. The 60% and 70% cut points are truncated to integer indices 26,294 and 30,676. Each 48-hour input and its complete 72-hour target remain within one split, without borrowing history from the preceding split. Counts below precede any training or adapter-fitting subsampling. Overlapping forecasts are evaluated as separate origin–lead instances.

Table 7: Shared temporal splits for France, Europe, and Global. Start times are inclusive and end times exclusive; all times are UTC. Window counts use a one-hour stride.

Split	Start	End	Valid windows
Training	2017-01-01 00:00	2020-01-01 14:00	26,175
Validation	2020-01-01 14:00	2020-07-02 04:00	4,263
Test	2020-07-02 04:00	2022-01-01 00:00	13,029

A.4 Spatial Subset Definitions and Additional Statistics

Global contains all 2,504 M^2-CS stations. Europe is defined using a frozen World Geodetic System 1984 (WGS84) polygon bounded by $32°$W–$70°$E and $34°$N–$82°$N, based on the physical European region and supplemented with country boundaries from the Geographical Information System of the Commission (GISCO). Transcontinental countries are included only through their intersection with the European polygon. The resulting subset contains 678 M^2-CS and 305 M^2-RA stations. France is the subset assigned the GISCO country code `FR`, containing 171 M^2-CS and 5 M^2-RA stations. Spatial membership does not modify the quality tier of a station.

The median nearest-neighbor distances are 42.97 km, 32.62 km, and 24.80 km for Global, Europe, and France, respectively. Their corresponding P_{10}–P_{90} ranges are 8.22–122.44 km, 5.79–83.03 km, and 1.07–54.04 km, demonstrating progressively denser station configurations across the three settings.

B Benchmark Implementation Details

B.1 Model Hyperparameters across Datasets

To ensure a consistent training comparison, all trained or fine-tuned backbones use Adam, MSE as the forecasting loss, an initial learning rate of 10^{-4}, and the type1 learning-rate decay schedule across France, Europe, and Global. Model-specific auxiliary objectives and regularization are retained where applicable and are described below. Timer, TimeMoE, and Moirai are evaluated in zero-shot mode and therefore have no local optimization settings.

Batch-size triples below follow the order France/Europe/Global and count temporal windows before wrapper reshaping; each window contains all stations. The listed batch sizes are the actual values used in the experiments. Spatial models use 176, 983, and 2,504 stations, respectively, together with the corresponding coordinates. Training lasts at most 30 epochs, with early-stopping patience 10/5/5. Training-window strides are 1/6/6, while validation and testing use stride 1, so training budgets differ across spatial scales. Architecture settings are shared across datasets unless stated otherwise. Weight decay is zero and gradient clipping is disabled unless specified below. Global automatic mixed-precision (AMP) settings are reported per model, with cross-region differences stated where applicable. Reused deterministic predictions are not treated as independent training runs.

DLinear. We use a moving-average window of 25 and shared seasonal and trend linear mappings from 48 to 72 steps (`individual=False`). Batch sizes are $(2048, 384, 864)$. AMP is disabled on Global.

xPatch. We use patch length 16, stride 8, replicated end padding, EMA trend coefficient 0.3, and RevIN. Seasonal and trend streams process each station–variable series independently and restore the instance scale at output. Batch sizes are $(1024, 192, 128)$; Global AMP is enabled.

PatchTST. Patch length and stride are 16 and 8. The encoder has 3 layers, model dimension 128, 16 attention heads, feed-forward dimension 256, and dropout 0.2. Station–variable series are processed independently with instance normalization and output rescaling. Batch sizes are $(128, 24, 8)$; Global AMP is disabled.

Timer. We use `thuml/timer-base-84m` for zero-shot univariate forecasting, processing at most 4,096 series per internal chunk. Short contexts are padded to a token, and window normalization uses a safe standard deviation for constant inputs. The model generates 72 steps autoregressively. Inference batch sizes are $(128, 24, 256)$. No local optimizer, training loss, or early stopping is used, and Global AMP is disabled. Global results for the latter two seeds reuse deterministic predictions.

TimeMoE. We use `Maple728/TimeMoE-50M` for zero-shot univariate forecasting, with internal chunks of 1,024 series. Each window is normalized by its own mean and standard deviation; 72 new steps are generated and inverse-transformed. Inference batch sizes are $(4, 12, 12)$. There is no

local training, and Global AMP is disabled. Global results for the latter two seeds reuse deterministic predictions.

iTransformer. We use 4 encoder layers, model dimension 256, 8 heads, feed-forward dimension 512, factor 3, and dropout 0.1. Each station's four variables are modeled through inverted variable tokens, with instance normalization and output rescaling. Batch sizes are $(128, 24, 16)$; Global AMP is disabled.

TQNet. We use a 24-hour cycle, the MLP variant, RevIN, dropout 0.5, learnable periodic queries, and 4-head variable attention. The cycle phase is the forecast-start index modulo 24; time-marker column 2 is only a fallback. Batch sizes are $(1024, 192, 144)$; Global AMP is disabled.

DUET. We use model dimension 256, 8 heads, 2 encoder layers, feed-forward dimension 512, hidden size 128, and 4 experts with top-1 routing. Both `CI` and noisy gating are enabled; dropout and fully connected dropout are 0.2. Each station's four variables are jointly supplied to the model. The forecasting objective is augmented with an expert load-balancing loss of weight 1. Batch sizes are $(256, 48, 48)$; Global AMP is disabled.

Moirai. We use `Salesforce/moirai-1.0-R-small`, with context length 48 and prediction length 72. At each station, the four target variables are jointly supplied in one inference call, allowing Any-variate Attention to model cross-variable dependencies. Point predictions are obtained by averaging over the sample dimension of the probabilistic forecasts. Inference batch sizes are $(8, 16, 8)$. There is no local training, and Global AMP is disabled.

Timer-XL. We load the pretrained Timer-XL checkpoint and fine-tune all parameters. The effective checkpoint architecture has model dimension 1,024, 8 heads, 8 layers, feed-forward dimension 2,048, and input/output tokens of length 96. Normalization is enabled; covariates and flash attention are disabled. History is left-padded to a 96-step token, and the first 72 steps of the final output token are retained. Gradient norms are clipped at 5. Batch sizes are $(64, 12, 4)$; Global AMP is disabled. Checkpoint dimensions override generic parser defaults.

Corrformer. We use 2 encoder layers, 1 decoder layer, model dimension 256, 8 heads, feed-forward dimension 512, moving-average window 25, temporal and spatial factors of 1, and one decoder TCN layer. Decoder context length is 24. Station hierarchies are $[11, 16]$, $[983]$, and $[8, 313]$ for France, Europe, and Global. Batch sizes are $(8, 1, 1)$. Global uses FP32 with saved backward activations offloaded to CPU to reduce GPU memory use.

S^{2}Transformer. We use model dimension 256, 8 heads, 2 encoder layers, feed-forward dimension 512, and dropout 0.1. K-means clustering of regional coordinates constructs geographic partitions with 16 and 4 groups, whose memberships vary by dataset. Variables are processed separately. Weight decay is 10^{-4}, and gradient norms are clipped at 5. Batch sizes are $(384, 64, 24)$; Global AMP is enabled.

CDPNet. We use feature dimension 32, a 32×32 grid, and 10 spatial neighbors. Regional coordinates determine the grid extent and station-to-grid interpolation; variables share the spatial network. Weight decay is 10^{-4}. The first epoch supervises future targets directly; subsequent epochs use a shifted sequence formed from concatenated history and targets, and checkpoint selection starts in the second epoch. Batch sizes are $(256, 48, 16)$; Global AMP is disabled.

EasyST. History, station, time, and transition embeddings each have dimension 64, and the student has 3 MLP layers. A frozen STELLA checkpoint from the same dataset and seed supplies the teacher. Gradient norms are clipped at 5. The forecasting objective additionally includes a teacher-error-gated prediction term weighted by 0.3 (`teacher_delta=0.1`) and latent KL regularization weighted by 0.001. Validation excludes the teacher and KL terms. Batch sizes are $(512, 96, 32)$; Global AMP is enabled.

STELLA. We use model dimension 32, 2 layers, dropout 0.2, coordinate-based positional encoding (`no_if_rel`), and residual connections. Variables are folded into the batch dimension and share

the spatial network. Weight decay is 5×10^{-4}, and gradient norms are clipped at 5. Batch sizes are $(1024, 192, 512)$; Global AMP is disabled.

HiSTGNN. We use four input features, graph-convolution depth 2, convolution channels 32, propagation coefficient 0.05, dropout 0.3, and 3 internal hierarchical layers with station and variable graph convolutions. Weight decay is 10^{-4}, and gradient norms are clipped at 5. Batch sizes are $(64, 20, 10)$. AMP is disabled on France and Europe and enabled on Global, where the evaluation batch size is 21.

B.2 Adapter Fitting and Parameter Efficiency

Feature construction. For a fixed forecast step and target variable (h, v), each training origin–station pair (t, n) contributes one row to $\mathbf{\Phi}_{h,v}$. With both interaction branches enabled, the feature and coefficient vectors are

$$\begin{aligned} \boldsymbol{\phi}_{t,n}^{(h,v)} &= \left[P_{t,h,n,v},\ 1,\ (\widetilde{P}_{t,h,n,u})_{u\neq v},\ \delta_{t,h,n,v}^{(1)},\ \delta_{t,h,n,v}^{(2)}\right]^\top, \\ \boldsymbol{\beta}_{h,v} &= \left[a_{h,v},\ b_{h,v},\ (W_{u,v}^{(h)})_{u\neq v},\ \gamma_{h,v}^{(1)},\ \gamma_{h,v}^{(2)}\right]^\top, \end{aligned} \tag{19}$$

where $\delta_{t,h,n,v}^{(r)} = (\mathbf{A}^r \mathbf{Z}_{t,h,:,v} - \mathbf{Z}_{t,h,:,v})_n$. The first feature is the original backbone prediction for calibration; the cross-variable block uses predictions normalized across variables, following Equation 4. The station features use the station-normalized predictions from Equation 6. All features are fixed before fitting. For an inactive branch, we omit its features and coefficients. Stacking the row vectors $(\boldsymbol{\phi}_{t,n}^{(h,v)})^\top$ gives $\mathbf{\Phi}_{h,v} \in \mathbb{R}^{M\times d}$, while the corresponding targets are $r_{t,n}^{(h,v)} = Y_{t,h,n,v} - P_{t,h,n,v}$. Thus, $d = V + 3$ for both branches, $d = V + 1$ for the variable branch, and $d = 4$ for the station branch, each including two calibration coefficients.

Branch normalization and geographic graph. Both interaction branches use population variance and the transform $(x - \mathrm{mean}(x))/\sqrt{\mathrm{var}(x) + 10^{-5}}$, without learnable affine parameters. The variable branch normalizes over the V predictions at fixed (t, h, n) to obtain $\widetilde{\mathbf{P}}_{t,h,n,:}$, whereas the station branch normalizes over the N stations at fixed (t, h, v) to obtain $\mathbf{Z}$. The calibration feature remains the original dataset-standardized prediction. All three branches read the same frozen-backbone output in parallel. Self-variable features are omitted from the regression, enforcing zero diagonal variable-interaction weights.

The fixed eight-neighbor graph uses latitude and longitude only. Distances d_{nj} are haversine central angles in radians. For each station, τ_n is the fourth-smallest of its eight neighbor distances (the lower median), bounded below by 10^{-8} radians. Self-neighbors are excluded, while distinct stations with identical coordinates are retained. Row-wise softmax of $-d_{nj}/\tau_n$ gives $\mathbf{A}$, which need not be symmetric. The two-hop feature is computed by applying this same fixed graph twice, $\mathbf{A}(\mathbf{AZ}) - \mathbf{Z}$.

Analytical solution and batch-wise accumulation. Differentiating the quadratic objective in Equation 7 yields the normal equations and, when the coefficient matrix is nonsingular, the corresponding inverse expression:

$$\begin{aligned} (\mathbf{\Phi}_{h,v}^\top \mathbf{\Phi}_{h,v} + M\lambda \mathbf{D})\boldsymbol{\beta}_{h,v}^\star &= \mathbf{\Phi}_{h,v}^\top \mathbf{r}_{h,v}, \\ \boldsymbol{\beta}_{h,v}^\star &= (\mathbf{\Phi}_{h,v}^\top \mathbf{\Phi}_{h,v} + M\lambda \mathbf{D})^{-1} \mathbf{\Phi}_{h,v}^\top \mathbf{r}_{h,v}. \end{aligned} \tag{20}$$

With the default calibration branch enabled, $\mathbf{D} = \mathrm{diag}(1, 0, 1, \ldots, 1)$ leaves only the intercept unpenalized. The candidate set includes $\lambda = 0$, so invertibility is not assumed for every candidate. In implementation, we solve the small normal-equation system in CPU float64 using a linear solver, falling back to a least-squares solve if the direct solve raises a linear-algebra error. We never explicitly form the inverse.

The full feature matrix need not be stored. For training batch b, let $\mathbf{\Phi}_{h,v}^{(b)}$ and $\mathbf{r}_{h,v}^{(b)}$ contain its feature rows and residuals. We accumulate only

$$\mathbf{G}_{h,v} = \sum_b (\mathbf{\Phi}_{h,v}^{(b)})^\top \mathbf{\Phi}_{h,v}^{(b)}, \qquad \mathbf{q}_{h,v} = \sum_b (\mathbf{\Phi}_{h,v}^{(b)})^\top \mathbf{r}_{h,v}^{(b)}, \tag{21}$$

and the total sample count M. After all batches have been processed, fitting reduces to $(\mathbf{G}_{h,v} + M\lambda\mathbf{D})\boldsymbol{\beta}^{\star}_{h,v} = \mathbf{q}_{h,v}$. These are exactly the full-data normal equations, up to floating-point arithmetic, and require one small solve per (h, v) for a fixed λ. Candidate regularization strengths reuse the same accumulated statistics and are selected by overall standardized validation MSE; no additional training-data pass is required. Excluding prediction and feature generation, accumulation costs $O(Md^2)$ and the final solve costs $O(d^3)$ per (h, v), with only $O(d^2)$ persistent storage. Since $V = 4$, each system has at most seven coefficients. This avoids iterative gradient updates, backpropagation, and optimizer-state storage.

Regularization selection and numerical implementation. We select a single λ per adapter run from the ordered set $\{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Coefficients are solved separately for each (h, v), but the selected λ is shared across all forecast steps, variables, and enabled branches. Selection minimizes aggregate validation MSE over all evaluated origins, stations, steps, and variables; exact ties select the smaller candidate, without a tolerance. The final coefficients use training statistics only, with no training–validation refit.

Parameter efficiency. Table 8 reports parameter counts for representative baselines and the recorded adapter configurations at seed 2024. Counts include independent coefficients that participate in prediction, excluding buffers, duplicate references to shared parameters, and masked adapter coefficients. Adapter counts include calibration but exclude the frozen backbone.

Table 8: Parameter counts across spatial scales. The upper block lists backbone parameters; the lower block lists only additional adapter parameters for the recorded branch configurations.

Paradigm / branch	Model	Global	Europe	France
Backbone parameters				
SSSV	xPatch	71,076	71,076	71,076
SSSV	PatchTST	455,112	455,112	455,112
SSMV	iTransformer	2,139,976	2,139,976	2,139,976
SSMV	TQNet	596,840	596,840	596,840
MSSV	S^2Transformer	2,404,936	2,404,936	2,404,936
MSSV	Corrformer	144,501,772	144,501,772	144,501,772
Additional adapter parameters				
Both	xPatch	2,016	2,016	2,016
Both	PatchTST	2,016	2,016	2,016
Variable	S^2Transformer	1,440	1,440	1,440
Variable	Corrformer	1,440	1,440	1,440
Station	iTransformer	1,152	1,152	1,152
Station	TQNet	1,152	1,152	1,152

Calibration contributes $2HV$ coefficients, variable interactions contribute $HV(V-1)$, and the two-hop station branch contributes $2HV$. With $H = 72$ and $V = 4$, station-only, variable-only, and joint adapters therefore contain 1,152, 1,440, and 2,016 coefficients, respectively. These counts are independent of the number of stations, so they remain unchanged across France ($N = 176$), Europe ($N = 983$), and Global ($N = 2{,}504$). For iTransformer and TQNet, the station-only adapter adds just 1,152 parameters, or 0.054% and 0.193% of their respective backbone sizes. The joint adapter adds 2.84% of xPatch's parameters and 0.44% of PatchTST's; the variable-only adapter adds 0.060% of S^2Transformer's and 0.0010% of Corrformer's. This station-shared parameterization supports interaction modeling with minimal additional model size.

C Additional Results

C.1 Reporting Conventions and Adapter Configurations

All comparisons use 48 hours of observations to predict 72 hours. The following tables report per-variable MSE and MAE. Entries are means $\pm$ sample standard deviations across the recorded seeds 2024, 2025, and 2026 (ddof $= 1$). Standard deviations describe variation between run-level

metrics, not the distribution of individual forecast errors. Values are rounded to four decimals; 0.0000 can denote a smaller nonzero deviation. Timer, TimeMoE, and Moirai use deterministic pretrained inference, so repeated values do not represent independently trained backbones.

For both MSE and MAE, $\Delta = 100(E_{\text{adapt}}/E_{\text{base}} - 1)$ denotes relative change: negative values indicate improvement and positive values indicate deterioration. Percentages are calculated before rounding. Group means weight models equally; group-level changes compare these means rather than averaging model-wise percentages. HiSTGNN is always shown without adaptation and is excluded from adapter improvement counts.

Evaluated branches. Model groups follow Table 2. DLinear, xPatch, PatchTST, Timer, and TimeMoE use both interaction branches; iTransformer, TQNet, DUET, Moirai, and Timer-XL use the station branch; Corrformer, S^{2}Transformer, CDPNet, EasyST, and STELLA use the variable branch. Full adapters additionally include calibration.

C.2 Seed-Level Significance Tests

We assess whether the MSE reductions in Table 3 are consistent across random seeds. For each stochastic backbone and dataset, we apply a two-sided paired t-test to the overall standardized MSE from seeds 2024, 2025, and 2026, pairing the native and adapted results from the same seed. The null hypothesis is that the mean paired difference is zero. We then control the family-wise error rate within each dataset using Holm correction over its 12 valid comparisons, with $\alpha = 0.05$. Timer, TimeMoE, and Moirai are excluded because their deterministic pretrained predictions are reused across seeds, while HiSTGNN has no adapted counterpart.

Table 9: Seed-level significance tests for the MSE reductions in Table 3. Entries are two-sided paired t-test p-values after Holm correction within each dataset; bold values are significant at $\alpha = 0.05$. Each test uses three paired seeds ($df = 2$).

Paradigm	Model	France	Europe	Global
SSSV	DLinear	$\mathbf{< 0.001}$	$\mathbf{< 0.001}$	$\mathbf{0.0044}$
	xPatch	$\mathbf{< 0.001}$	$\mathbf{< 0.001}$	$\mathbf{0.0058}$
	PatchTST	$\mathbf{< 0.001}$	$\mathbf{< 0.001}$	$\mathbf{< 0.001}$
SSMV	iTransformer	$\mathbf{< 0.001}$	$\mathbf{< 0.001}$	$\mathbf{0.0058}$
	TQNet	$\mathbf{< 0.001}$	$\mathbf{0.0069}$	$\mathbf{0.0313}$
	DUET	$\mathbf{0.0043}$	$\mathbf{0.0193}$	$\mathbf{0.0128}$
	Timer-XL	$\mathbf{< 0.001}$	$\mathbf{0.0069}$	$\mathbf{< 0.001}$
MSSV	Corrformer	$\mathbf{0.0043}$	$\mathbf{0.0045}$	$\mathbf{0.0035}$
	S^{2}Transformer	$\mathbf{< 0.001}$	$\mathbf{0.0045}$	$\mathbf{0.0042}$
	CDPNet	0.0808	$\mathbf{0.0193}$	$\mathbf{0.0313}$
	EasyST	$\mathbf{< 0.001}$	$\mathbf{0.0013}$	$\mathbf{0.0038}$
	STELLA	$\mathbf{< 0.001}$	$\mathbf{0.0045}$	0.0547

After within-dataset correction, 11 of 12 testable backbones are significant on France, all 12 are significant on Europe, and 11 of 12 are significant on Global. The exceptions are CDPNet on France ($p_{\text{Holm}} = 0.0808$) and STELLA on Global ($p_{\text{Holm}} = 0.0547$). If all 36 tests across the three datasets are treated as a single Holm family, 30 remain significant; Europe DUET and CDPNet and Global TQNet and CDPNet additionally exceed 0.05.

C.3 Complete Per-Variable MSE and MAE

Tables 10–21 provide one table for every dataset–variable pair. All 180 adapted model–dataset–variable combinations improve MSE. At the displayed precision of the relative changes, MAE decreases in 56/60, 54/60, and 55/60 combinations on France, Europe, and Global, respectively. One additional Europe comparison (TQNet on station pressure) rounds to 0.00% and is reported as unchanged at this precision. The remaining 4, 5, and 5 comparisons show MAE increases; the adapter is fitted and selected using squared error.

Table 10: **France — Air temperature.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.3146±0.0003	0.2928±0.0000	−6.94	0.4411±0.0005	0.4259±0.0000	−3.44
xPatch	0.1977±0.0004	0.1921±0.0003	−2.81	0.3371±0.0003	0.3342±0.0003	−0.89
PatchTST	0.1987±0.0005	0.1861±0.0005	−6.36	0.3393±0.0005	0.3298±0.0005	−2.79
Timer	0.3078±0.0000	0.2718±0.0000	−11.67	0.4235±0.0000	0.4034±0.0000	−4.74
TimeMoE	0.2126±0.0000	0.1976±0.0000	−7.07	0.3476±0.0000	0.3377±0.0000	−2.83
iTransformer	0.1693±0.0007	0.1639±0.0008	−3.18	0.3099±0.0007	0.3061±0.0007	−1.21
TQNet	0.1881±0.0006	0.1822±0.0006	−3.12	0.3271±0.0006	0.3239±0.0005	−0.98
DUET	0.2082±0.0065	0.2008±0.0040	−3.56	0.3473±0.0053	0.3428±0.0040	−1.31
Moirai	0.2326±0.0000	0.1992±0.0000	−14.36	0.3602±0.0000	0.3393±0.0000	−5.80
Timer-XL	0.1687±0.0003	0.1620±0.0002	−3.97	0.3087±0.0002	0.3038±0.0002	−1.58
Corrformer	0.1787±0.0043	0.1658±0.0067	−7.21	0.3202±0.0030	0.3103±0.0053	−3.12
S^2Transformer	0.1773±0.0011	0.1720±0.0009	−2.95	0.3241±0.0010	0.3194±0.0009	−1.44
CDPNet	0.1636±0.0068	0.1577±0.0064	−3.61	0.3101±0.0057	0.3042±0.0057	−1.88
EasyST	0.1783±0.0008	0.1743±0.0004	−2.25	0.3221±0.0007	0.3185±0.0003	−1.12
STELLA	0.1987±0.0006	0.1915±0.0003	−3.62	0.3421±0.0009	0.3360±0.0006	−1.77
HiSTGNN	0.1438±0.0063	–	–	0.2883±0.0062	–	–

Table 11: **France — Wind speed.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.8359±0.0003	0.7586±0.0000	−9.25	0.6746±0.0014	0.6506±0.0000	−3.56
xPatch	0.7956±0.0006	0.7054±0.0003	−11.34	0.6394±0.0003	0.6171±0.0003	−3.49
PatchTST	0.7811±0.0002	0.7029±0.0001	−10.01	0.6374±0.0004	0.6165±0.0001	−3.28
Timer	0.9058±0.0000	0.7490±0.0000	−17.31	0.6968±0.0000	0.6434±0.0000	−7.68
TimeMoE	0.8696±0.0000	0.7362±0.0000	−15.34	0.6694±0.0000	0.6328±0.0000	−5.47
iTransformer	0.7184±0.0003	0.6811±0.0007	−5.20	0.6115±0.0006	0.6140±0.0002	+0.42
TQNet	0.7514±0.0009	0.6957±0.0005	−7.41	0.6216±0.0004	0.6120±0.0002	−1.55
DUET	0.7989±0.0085	0.7180±0.0046	−10.12	0.6386±0.0023	0.6244±0.0021	−2.23
Moirai	0.9476±0.0000	0.7383±0.0000	−22.09	0.6899±0.0000	0.6318±0.0000	−8.42
Timer-XL	0.7136±0.0005	0.6838±0.0009	−4.17	0.6100±0.0011	0.6022±0.0002	−1.28
Corrformer	0.7506±0.0100	0.6931±0.0071	−7.67	0.6259±0.0089	0.6042±0.0016	−3.47
S^2Transformer	0.6860±0.0002	0.6820±0.0002	−0.59	0.5990±0.0002	0.5970±0.0001	−0.33
CDPNet	0.6636±0.0109	0.6486±0.0169	−2.26	0.5818±0.0087	0.5855±0.0088	+0.65
EasyST	0.6740±0.0009	0.6697±0.0009	−0.64	0.5996±0.0006	0.5986±0.0005	−0.16
STELLA	0.7637±0.0010	0.7150±0.0018	−6.38	0.6324±0.0004	0.6187±0.0009	−2.18
HiSTGNN	0.6273±0.0052	–	–	0.5621±0.0041	–	–

Table 12: **France — Relative humidity.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.7059±0.0007	0.6835±0.0000	−3.16	0.6742±0.0020	0.6734±0.0000	−0.12
xPatch	0.4959±0.0007	0.4671±0.0006	−5.82	0.5305±0.0005	0.5316±0.0005	+0.22
PatchTST	0.4882±0.0005	0.4657±0.0005	−4.61	0.5346±0.0001	0.5311±0.0003	−0.67
Timer	0.7680±0.0000	0.6644±0.0000	−13.50	0.6638±0.0000	0.6551±0.0000	−1.31
TimeMoE	0.5592±0.0000	0.5141±0.0000	−8.07	0.5644±0.0000	0.5635±0.0000	−0.16
iTransformer	0.4309±0.0005	0.4111±0.0005	−4.58	0.4899±0.0004	0.4887±0.0003	−0.24
TQNet	0.4541±0.0005	0.4341±0.0004	−4.41	0.5065±0.0003	0.5043±0.0003	−0.42
DUET	0.5122±0.0132	0.4826±0.0133	−5.78	0.5417±0.0089	0.5402±0.0094	−0.28
Moirai	0.5998±0.0000	0.5006±0.0000	−16.55	0.5707±0.0000	0.5540±0.0000	−2.93
Timer-XL	0.4323±0.0004	0.4114±0.0005	−4.84	0.4921±0.0008	0.4907±0.0003	−0.28
Corrformer	0.4527±0.0006	0.4202±0.0038	−7.20	0.5015±0.0008	0.4958±0.0023	−1.14
S^2Transformer	0.4386±0.0007	0.4336±0.0008	−1.14	0.5130±0.0005	0.5090±0.0007	−0.77
CDPNet	0.4252±0.0080	0.4218±0.0107	−0.80	0.4926±0.0064	0.4910±0.0049	−0.33
EasyST	0.4473±0.0016	0.4393±0.0016	−1.80	0.5189±0.0009	0.5117±0.0011	−1.39
STELLA	0.4909±0.0015	0.4736±0.0016	−3.52	0.5357±0.0011	0.5370±0.0011	+0.24
HiSTGNN	0.3892±0.0054	–	–	0.4662±0.0035	–	–

Table 13: **France — Station pressure.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.1346±0.0004	0.1069±0.0000	−20.53	0.2715±0.0013	0.2379±0.0000	−12.38
xPatch	0.1005±0.0004	0.0922±0.0003	−8.25	0.2211±0.0006	0.2141±0.0006	−3.16
PatchTST	0.0941±0.0006	0.0856±0.0005	−8.99	0.2121±0.0009	0.2044±0.0008	−3.64
Timer	0.1030±0.0000	0.0929±0.0000	−9.73	0.2222±0.0000	0.2133±0.0000	−4.03
TimeMoE	0.1002±0.0000	0.0901±0.0000	−10.06	0.2174±0.0000	0.2084±0.0000	−4.13
iTransformer	0.0824±0.0003	0.0777±0.0003	−5.69	0.1934±0.0002	0.1892±0.0002	−2.18
TQNet	0.0881±0.0003	0.0818±0.0002	−7.19	0.2023±0.0003	0.1965±0.0004	−2.83
DUET	0.1014±0.0035	0.0933±0.0031	−7.99	0.2221±0.0041	0.2151±0.0041	−3.14
Moirai	0.1067±0.0000	0.0887±0.0000	−16.90	0.2223±0.0000	0.2062±0.0000	−7.23
Timer-XL	0.0786±0.0003	0.0753±0.0003	−4.26	0.1880±0.0005	0.1849±0.0004	−1.63
Corrformer	0.0916±0.0033	0.0879±0.0027	−3.98	0.2042±0.0039	0.2007±0.0029	−1.69
S^2Transformer	0.2087±0.0043	0.1725±0.0030	−17.31	0.2338±0.0010	0.2274±0.0014	−2.75
CDPNet	0.0721±0.0009	0.0711±0.0005	−1.39	0.1851±0.0040	0.1816±0.0010	−1.88
EasyST	0.0773±0.0003	0.0761±0.0003	−1.49	0.1906±0.0003	0.1885±0.0005	−1.09
STELLA	0.0978±0.0007	0.0959±0.0005	−1.92	0.2238±0.0018	0.2211±0.0016	−1.22
HiSTGNN	0.0712±0.0021	–	–	0.1812±0.0021	–	–

Table 14: **Europe — Air temperature.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.2270±0.0003	0.2107±0.0000	−7.15	0.3684±0.0004	0.3536±0.0000	−4.03
xPatch	0.1481±0.0002	0.1445±0.0001	−2.41	0.2861±0.0002	0.2836±0.0002	−0.89
PatchTST	0.1505±0.0006	0.1432±0.0006	−4.84	0.2898±0.0008	0.2836±0.0008	−2.16
Timer	0.2039±0.0000	0.1859±0.0000	−8.83	0.3386±0.0000	0.3263±0.0000	−3.63
TimeMoE	0.1561±0.0000	0.1482±0.0000	−5.09	0.2917±0.0000	0.2858±0.0000	−2.03
iTransformer	0.1362±0.0010	0.1328±0.0010	−2.49	0.2730±0.0012	0.2702±0.0011	−1.02
TQNet	0.1503±0.0030	0.1466±0.0030	−2.41	0.2890±0.0035	0.2864±0.0034	−0.92
DUET	0.1779±0.0381	0.1739±0.0381	−2.29	0.3180±0.0392	0.3147±0.0388	−1.02
Moirai	0.1685±0.0000	0.1502±0.0000	−10.86	0.2998±0.0000	0.2875±0.0000	−4.09
Timer-XL	0.1266±0.0004	0.1235±0.0001	−2.47	0.2615±0.0004	0.2589±0.0001	−1.00
Corrformer	0.1237±0.0008	0.1210±0.0007	−2.21	0.2593±0.0008	0.2574±0.0007	−0.74
S^2Transformer	0.1357±0.0008	0.1346±0.0007	−0.84	0.2785±0.0011	0.2772±0.0010	−0.45
CDPNet	0.1156±0.0025	0.1148±0.0022	−0.69	0.2557±0.0011	0.2548±0.0008	−0.35
EasyST	0.1371±0.0001	0.1349±0.0000	−1.56	0.2766±0.0001	0.2747±0.0001	−0.70
STELLA	0.1754±0.0060	0.1685±0.0052	−3.92	0.3211±0.0061	0.3146±0.0054	−2.02
HiSTGNN	0.1231±0.0093	–	–	0.2622±0.0090	–	–

Table 15: **Europe — Wind speed.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.7531±0.0004	0.6782±0.0000	−9.94	0.6265±0.0014	0.6085±0.0000	−2.87
xPatch	0.6890±0.0004	0.6318±0.0002	−8.29	0.5836±0.0002	0.5780±0.0002	−0.96
PatchTST	0.6775±0.0008	0.6272±0.0001	−7.42	0.5830±0.0002	0.5762±0.0001	−1.18
Timer	0.7625±0.0000	0.6617±0.0000	−13.23	0.6238±0.0000	0.5979±0.0000	−4.16
TimeMoE	0.7422±0.0000	0.6554±0.0000	−11.68	0.6048±0.0000	0.5905±0.0000	−2.35
iTransformer	0.6455±0.0006	0.6171±0.0007	−4.40	0.5682±0.0008	0.5680±0.0005	−0.05
TQNet	0.6807±0.0035	0.6294±0.0005	−7.53	0.5838±0.0025	0.5799±0.0022	−0.66
DUET	0.7174±0.0349	0.6588±0.0256	−8.17	0.5938±0.0182	0.5901±0.0164	−0.63
Moirai	0.8183±0.0000	0.6647±0.0000	−18.78	0.6253±0.0000	0.5948±0.0000	−4.88
Timer-XL	0.6281±0.0003	0.6086±0.0004	−3.11	0.5599±0.0015	0.5601±0.0002	+0.03
Corrformer	0.6263±0.0019	0.6035±0.0011	−3.64	0.5598±0.0011	0.5594±0.0006	−0.08
S^2Transformer	0.6162±0.0006	0.6134±0.0005	−0.45	0.5661±0.0004	0.5657±0.0003	−0.07
CDPNet	0.5875±0.0007	0.5722±0.0032	−2.62	0.5358±0.0018	0.5430±0.0023	+1.34
EasyST	0.5909±0.0004	0.5899±0.0005	−0.16	0.5539±0.0006	0.5518±0.0003	−0.37
STELLA	0.7013±0.0031	0.6557±0.0037	−6.51	0.5997±0.0050	0.5874±0.0085	−2.06
HiSTGNN	0.5662±0.0105	–	–	0.5264±0.0047	–	–

Table 16: **Europe — Relative humidity.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.7719±0.0006	0.7191±0.0000	−6.85	0.6820±0.0019	0.6794±0.0000	−0.39
xPatch	0.5427±0.0001	0.5119±0.0002	−5.68	0.5439±0.0003	0.5416±0.0003	−0.43
PatchTST	0.5371±0.0015	0.5131±0.0016	−4.46	0.5502±0.0015	0.5499±0.0011	−0.05
Timer	0.7665±0.0000	0.6671±0.0000	−12.98	0.6468±0.0000	0.6440±0.0000	−0.43
TimeMoE	0.5974±0.0000	0.5476±0.0000	−8.33	0.5686±0.0000	0.5617±0.0000	−1.22
iTransformer	0.4993±0.0008	0.4768±0.0013	−4.51	0.5237±0.0007	0.5304±0.0011	+1.28
TQNet	0.5341±0.0081	0.5099±0.0063	−4.53	0.5439±0.0025	0.5410±0.0039	−0.54
DUET	0.6405±0.1267	0.5999±0.1177	−6.34	0.5975±0.0695	0.6044±0.0727	+1.15
Moirai	0.6382±0.0000	0.5370±0.0000	−15.86	0.5755±0.0000	0.5647±0.0000	−1.88
Timer-XL	0.4786±0.0017	0.4610±0.0004	−3.67	0.5116±0.0011	0.5102±0.0003	−0.26
Corrformer	0.4681±0.0009	0.4494±0.0008	−4.00	0.5017±0.0005	0.5079±0.0004	+1.24
S^2Transformer	0.4888±0.0007	0.4860±0.0008	−0.58	0.5317±0.0006	0.5314±0.0005	−0.07
CDPNet	0.4530±0.0008	0.4443±0.0025	−1.93	0.5019±0.0013	0.5009±0.0009	−0.18
EasyST	0.4943±0.0005	0.4889±0.0006	−1.09	0.5365±0.0002	0.5341±0.0004	−0.44
STELLA	0.5630±0.0046	0.5310±0.0047	−5.69	0.5649±0.0057	0.5643±0.0060	−0.11
HiSTGNN	0.4548±0.0210	–	–	0.5000±0.0092	–	–

Table 17: **Europe — Station pressure.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.0404±0.0004	0.0268±0.0000	−33.59	0.1505±0.0017	0.1128±0.0000	−25.07
xPatch	0.0238±0.0001	0.0236±0.0001	−0.64	0.1023±0.0003	0.1022±0.0003	−0.06
PatchTST	0.0236±0.0001	0.0229±0.0001	−2.99	0.1017±0.0003	0.1003±0.0003	−1.34
Timer	0.0240±0.0000	0.0232±0.0000	−3.04	0.1016±0.0000	0.1001±0.0000	−1.48
TimeMoE	0.0234±0.0000	0.0227±0.0000	−3.23	0.0986±0.0000	0.0972±0.0000	−1.47
iTransformer	0.0214±0.0000	0.0212±0.0000	−0.73	0.0937±0.0000	0.0936±0.0000	−0.10
TQNet	0.0236±0.0003	0.0235±0.0002	−0.52	0.1013±0.0010	0.1013±0.0009	0.00
DUET	0.0262±0.0015	0.0259±0.0014	−0.94	0.1089±0.0053	0.1085±0.0052	−0.29
Moirai	0.0245±0.0000	0.0234±0.0000	−4.55	0.0996±0.0000	0.0974±0.0000	−2.17
Timer-XL	0.0196±0.0000	0.0195±0.0000	−0.85	0.0882±0.0001	0.0880±0.0001	−0.26
Corrformer	0.0216±0.0003	0.0212±0.0003	−1.64	0.0941±0.0008	0.0936±0.0008	−0.54
S^2Transformer	0.0921±0.0031	0.0670±0.0017	−27.23	0.1414±0.0013	0.1405±0.0018	−0.62
CDPNet	0.0199±0.0006	0.0195±0.0004	−1.84	0.0919±0.0031	0.0900±0.0018	−2.13
EasyST	0.0209±0.0001	0.0206±0.0001	−1.42	0.0949±0.0004	0.0936±0.0005	−1.42
STELLA	0.0530±0.0055	0.0528±0.0055	−0.45	0.1743±0.0105	0.1737±0.0105	−0.33
HiSTGNN	0.0192±0.0008	–	–	0.0919±0.0029	–	–

Table 18: **Global — Air temperature.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.1837±0.0008	0.1685±0.0011	−8.25	0.3285±0.0009	0.3124±0.0012	−4.88
xPatch	0.1186±0.0010	0.1160±0.0011	−2.25	0.2507±0.0015	0.2495±0.0015	−0.46
PatchTST	0.1188±0.0000	0.1139±0.0001	−4.11	0.2514±0.0001	0.2478±0.0001	−1.45
Timer	0.1628±0.0000	0.1500±0.0000	−7.88	0.2979±0.0000	0.2895±0.0000	−2.80
TimeMoE	0.1258±0.0000	0.1208±0.0000	−3.95	0.2562±0.0000	0.2533±0.0000	−1.12
iTransformer	0.1019±0.0013	0.0999±0.0012	−1.89	0.2297±0.0017	0.2289±0.0017	−0.36
TQNet	0.1176±0.0017	0.1149±0.0019	−2.25	0.2493±0.0024	0.2481±0.0028	−0.47
DUET	0.1223±0.0022	0.1197±0.0019	−2.09	0.2559±0.0027	0.2548±0.0026	−0.45
Moirai	0.1398±0.0000	0.1262±0.0000	−9.74	0.2661±0.0000	0.2576±0.0000	−3.19
Timer-XL	0.0983±0.0001	0.0961±0.0001	−2.23	0.2250±0.0001	0.2236±0.0001	−0.60
Corrformer	0.0986±0.0003	0.0969±0.0004	−1.66	0.2272±0.0005	0.2267±0.0006	−0.22
S^2Transformer	0.0991±0.0008	0.0986±0.0007	−0.46	0.2337±0.0010	0.2333±0.0010	−0.21
CDPNet	0.0863±0.0007	0.0857±0.0008	−0.73	0.2183±0.0014	0.2174±0.0014	−0.39
EasyST	0.1072±0.0004	0.1067±0.0004	−0.44	0.2393±0.0007	0.2387±0.0006	−0.26
STELLA	0.1902±0.0519	0.1738±0.0401	−8.63	0.3342±0.0519	0.3192±0.0420	−4.49
HiSTGNN	0.0949±0.0029	–	–	0.2267±0.0037	–	–

Table 19: **Global — Wind speed.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.8466±0.0053	0.7770±0.0010	−8.22	0.6877±0.0024	0.6675±0.0006	−2.94
xPatch	0.7833±0.0041	0.7216±0.0003	−7.87	0.6473±0.0013	0.6348±0.0004	−1.94
PatchTST	0.7767±0.0005	0.7179±0.0002	−7.57	0.6473±0.0002	0.6333±0.0001	−2.17
Timer	0.8832±0.0000	0.7617±0.0000	−13.76	0.6968±0.0000	0.6576±0.0000	−5.62
TimeMoE	0.8533±0.0000	0.7477±0.0000	−12.38	0.6747±0.0000	0.6476±0.0000	−4.01
iTransformer	0.7134±0.0037	0.6843±0.0019	−4.08	0.6177±0.0018	0.6133±0.0013	−0.71
TQNet	0.7687±0.0163	0.7181±0.0036	−6.58	0.6431±0.0086	0.6343±0.0027	−1.37
DUET	0.7844±0.0109	0.7264±0.0044	−7.40	0.6451±0.0049	0.6474±0.0026	+0.36
Moirai	0.9503±0.0000	0.7581±0.0000	−20.22	0.7032±0.0000	0.6528±0.0000	−7.16
Timer-XL	0.7031±0.0001	0.6793±0.0002	−3.38	0.6134±0.0010	0.6104±0.0001	−0.48
Corrformer	0.7069±0.0007	0.6778±0.0003	−4.11	0.6166±0.0009	0.6096±0.0004	−1.13
S^2Transformer	0.6816±0.0005	0.6809±0.0005	−0.10	0.6127±0.0002	0.6123±0.0001	−0.06
CDPNet	0.6526±0.0043	0.6472±0.0059	−0.83	0.5899±0.0035	0.5941±0.0026	+0.70
EasyST	0.6597±0.0004	0.6572±0.0003	−0.38	0.6038±0.0002	0.6023±0.0001	−0.25
STELLA	0.8433±0.0545	0.7585±0.0257	−10.05	0.6836±0.0332	0.6527±0.0183	−4.53
HiSTGNN	0.6423±0.0059	–	–	0.5815±0.0013	–	–

Table 20: **Global — Relative humidity.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.7290±0.0071	0.6969±0.0040	−4.41	0.6677±0.0031	0.6673±0.0023	−0.05
xPatch	0.5100±0.0042	0.4878±0.0035	−4.36	0.5264±0.0023	0.5214±0.0024	−0.95
PatchTST	0.4980±0.0006	0.4829±0.0008	−3.03	0.5285±0.0009	0.5261±0.0006	−0.46
Timer	0.7494±0.0000	0.6601±0.0000	−11.92	0.6430±0.0000	0.6371±0.0000	−0.91
TimeMoE	0.5753±0.0000	0.5332±0.0000	−7.32	0.5577±0.0000	0.5611±0.0000	+0.61
iTransformer	0.4577±0.0031	0.4435±0.0025	−3.11	0.4972±0.0018	0.4903±0.0017	−1.38
TQNet	0.4930±0.0090	0.4771±0.0090	−3.23	0.5211±0.0064	0.5246±0.0073	+0.66
DUET	0.5236±0.0085	0.4961±0.0068	−5.25	0.5332±0.0053	0.5308±0.0050	−0.45
Moirai	0.6141±0.0000	0.5248±0.0000	−14.55	0.5638±0.0000	0.5555±0.0000	−1.47
Timer-XL	0.4505±0.0007	0.4371±0.0006	−2.97	0.4938±0.0006	0.4882±0.0004	−1.13
Corrformer	0.4496±0.0019	0.4367±0.0010	−2.87	0.4924±0.0010	0.4980±0.0008	+1.14
S^2Transformer	0.4467±0.0011	0.4462±0.0011	−0.11	0.5065±0.0005	0.5059±0.0006	−0.13
CDPNet	0.4382±0.0041	0.4349±0.0040	−0.76	0.4965±0.0023	0.4960±0.0023	−0.10
EasyST	0.4548±0.0004	0.4523±0.0003	−0.56	0.5110±0.0005	0.5089±0.0004	−0.41
STELLA	0.5912±0.0619	0.5540±0.0455	−6.29	0.5847±0.0390	0.5781±0.0278	−1.12
HiSTGNN	0.4291±0.0053	–	–	0.4855±0.0052	–	–

Table 21: **Global — Station pressure.** MSE and MAE (mean ± sample standard deviation). Δ (%) is signed relative change; negative is better.

	MSE ↓			MAE ↓		
Model	Base	+Adapter	Δ (%)	Base	+Adapter	Δ (%)
DLinear	0.0438±0.0013	0.0263±0.0000	−39.83	0.1603±0.0019	0.1163±0.0001	−27.50
xPatch	0.0230±0.0004	0.0229±0.0004	−0.59	0.1043±0.0012	0.1043±0.0012	−0.08
PatchTST	0.0224±0.0001	0.0218±0.0001	−2.71	0.1026±0.0004	0.1015±0.0004	−1.05
Timer	0.0232±0.0000	0.0225±0.0000	−2.85	0.1046±0.0000	0.1033±0.0000	−1.22
TimeMoE	0.0227±0.0000	0.0220±0.0000	−3.06	0.1021±0.0000	0.1009±0.0000	−1.16
iTransformer	0.0200±0.0003	0.0199±0.0003	−0.56	0.0943±0.0011	0.0933±0.0011	−1.00
TQNet	0.0230±0.0005	0.0229±0.0004	−0.47	0.1040±0.0015	0.1040±0.0014	−0.04
DUET	0.0258±0.0036	0.0256±0.0035	−0.74	0.1113±0.0081	0.1112±0.0083	−0.07
Moirai	0.0262±0.0000	0.0249±0.0000	−5.09	0.1074±0.0000	0.1053±0.0000	−1.95
Timer-XL	0.0190±0.0000	0.0189±0.0000	−0.43	0.0911±0.0001	0.0901±0.0000	−1.05
Corrformer	0.0205±0.0003	0.0202±0.0003	−1.30	0.0964±0.0009	0.0961±0.0008	−0.35
S^2Transformer	0.0358±0.0006	0.0334±0.0005	−6.67	0.1086±0.0007	0.1008±0.0007	−7.11
CDPNet	0.0168±0.0003	0.0165±0.0002	−1.86	0.0888±0.0010	0.0879±0.0009	−0.98
EasyST	0.0186±0.0001	0.0185±0.0001	−0.54	0.0933±0.0006	0.0926±0.0004	−0.67
STELLA	0.0894±0.0453	0.0862±0.0424	−3.60	0.2249±0.0669	0.2212±0.0632	−1.60
HiSTGNN	0.0179±0.0002	–	–	0.0927±0.0011	–	–

C.4 Individual Models within Each Paradigm

Figure 6 expands Figure 5 into all 16 backbones. The same model order and horizontal scale are used across datasets. Each connected pair compares a frozen backbone with its full adapter; HiSTGNN appears only as the native MSMV reference. Even within the same station–variable modeling paradigm, backbones benefit from interaction completion to different degrees. For example, among SSMV models on France, MSE decreases by 19.08% for Moirai but 4.36% for Timer-XL. Thus, the modeling paradigm alone does not determine the magnitude of the improvement; the benefit also varies with the backbone.

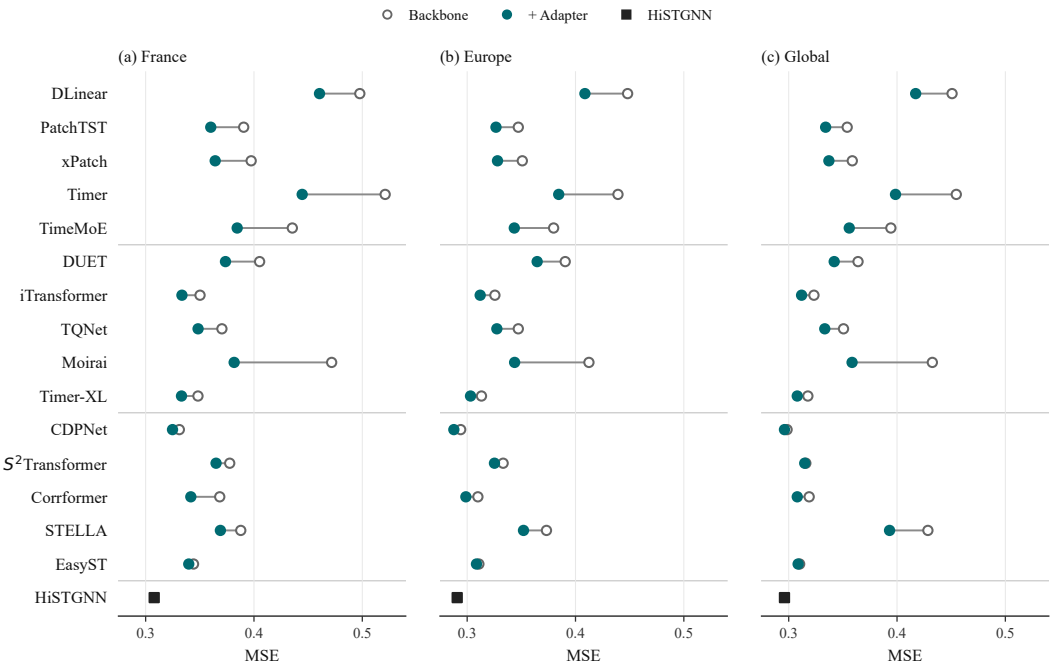

Figure 6: Individual-model counterpart of Figure 5. Open circles denote native predictions, filled circles denote adapted predictions, and squares mark the unadapted HiSTGNN. Thin rules separate SSSV, SSMV, MSSV, and MSMV; lower MSE is better.

C.5 Removing the Calibration Branch

The interaction-only variant fixes $a_{h,v} = b_{h,v} = 0$ and refits the same enabled interaction branches as the full adapter, with the backbone frozen. Table 22 reports MSE for France and Global. Interaction-only correction lowers MSE for all 15 backbones on both datasets, showing that useful gains persist without the explicit calibration term. The full adapter further improves every pair. For example, on Global, Moirai improves from 0.4326 to 0.3766 with interaction-only correction and to 0.3585 with the full adapter. These separately fitted variants establish complementary benefits without treating their improvements as an additive decomposition.

C.6 Sensitivity to Residual Fitting Sources

On France, we compare three residual fitting sources for PatchTST, iTransformer, and S^{2}Transformer over seeds 2024–2026. All variants share the same frozen backbone and adapter-validation set, with calibration and the corresponding interaction branches retained. *In-sample full* uses all 20,916 backbone-training forecast origins; *In-sample matched* uses its final contiguous 5,140 origins; and *Held-out* uses 5,140 origins from a subsequent block unseen during backbone training. Windows remain within disjoint temporal blocks, and checkpoint selection and adapter tuning use separate validation blocks. The original test period is unchanged; comparisons use the shared Base under this controlled split.

Table 22: Calibration ablation on France and Global. MSE is reported as mean ± sample standard deviation across seeds 2024–2026. Lower is better. **Full** includes calibration and the same interaction branches. HiSTGNN is excluded.

France			
Model	Base	Interaction only	Full
DLinear	0.4977 ± 0.0004	0.4800 ± 0.0004	**0.4605** ± 0.0000
xPatch	0.3974 ± 0.0004	0.3858 ± 0.0003	**0.3642** ± 0.0002
PatchTST	0.3905 ± 0.0004	0.3724 ± 0.0004	**0.3601** ± 0.0003
Timer	0.5211 ± 0.0000	0.4704 ± 0.0000	**0.4445** ± 0.0000
TimeMoE	0.4354 ± 0.0000	0.4028 ± 0.0000	**0.3845** ± 0.0000
iTransformer	0.3502 ± 0.0001	0.3423 ± 0.0000	**0.3335** ± 0.0001
TQNet	0.3704 ± 0.0004	0.3628 ± 0.0004	**0.3484** ± 0.0003
DUET	0.4052 ± 0.0074	0.3949 ± 0.0066	**0.3737** ± 0.0062
Moirai	0.4717 ± 0.0000	0.4094 ± 0.0000	**0.3817** ± 0.0000
Timer-XL	0.3483 ± 0.0001	0.3401 ± 0.0002	**0.3331** ± 0.0002
Corrformer	0.3684 ± 0.0031	0.3556 ± 0.0039	**0.3417** ± 0.0043
S^2Transformer	0.3776 ± 0.0015	0.3743 ± 0.0014	**0.3650** ± 0.0012
CDPNet	0.3311 ± 0.0057	0.3293 ± 0.0063	**0.3248** ± 0.0083
EasyST	0.3442 ± 0.0009	0.3428 ± 0.0009	**0.3398** ± 0.0008
STELLA	0.3878 ± 0.0005	0.3798 ± 0.0007	**0.3690** ± 0.0007
Global			
Model	Base	Interaction only	Full
DLinear	0.4508 ± 0.0030	0.4369 ± 0.0022	**0.4172** ± 0.0015
xPatch	0.3587 ± 0.0024	0.3478 ± 0.0019	**0.3371** ± 0.0012
PatchTST	0.3540 ± 0.0001	0.3414 ± 0.0002	**0.3341** ± 0.0002
Timer	0.4547 ± 0.0000	0.4154 ± 0.0000	**0.3986** ± 0.0000
TimeMoE	0.3943 ± 0.0000	0.3682 ± 0.0000	**0.3559** ± 0.0000
iTransformer	0.3233 ± 0.0021	0.3159 ± 0.0018	**0.3119** ± 0.0015
TQNet	0.3506 ± 0.0067	0.3417 ± 0.0053	**0.3333** ± 0.0036
DUET	0.3640 ± 0.0054	0.3543 ± 0.0046	**0.3420** ± 0.0032
Moirai	0.4326 ± 0.0000	0.3766 ± 0.0000	**0.3585** ± 0.0000
Timer-XL	0.3177 ± 0.0002	0.3110 ± 0.0001	**0.3079** ± 0.0002
Corrformer	0.3189 ± 0.0007	0.3138 ± 0.0006	**0.3079** ± 0.0004
S^2Transformer	0.3158 ± 0.0002	0.3154 ± 0.0002	**0.3148** ± 0.0002
CDPNet	0.2985 ± 0.0011	0.2975 ± 0.0013	**0.2961** ± 0.0016
EasyST	0.3101 ± 0.0003	0.3091 ± 0.0002	**0.3087** ± 0.0002
STELLA	0.4285 ± 0.0534	0.4220 ± 0.0540	**0.3931** ± 0.0384

Table 23: Sensitivity to residual fitting sources on France. MSE is reported as mean ± sample standard deviation over three seeds. Δ is the mean per-seed percentage change relative to the shared Base; negative values indicate improvement. Bold marks the lowest mean MSE per backbone.

Backbone	Residual source	MSE ↓	Δ (%)
PatchTST	Base	0.4030 ± 0.0001	—
	In-sample full	$\mathbf{0.3703} \pm 0.0001$	-8.11
	In-sample matched	0.3790 ± 0.0001	-5.94
	Held-out	0.3716 ± 0.0001	-7.79
iTransformer	Base	0.3813 ± 0.0005	—
	In-sample full	$\mathbf{0.3525} \pm 0.0003$	-7.54
	In-sample matched	0.3554 ± 0.0002	-6.79
	Held-out	0.3526 ± 0.0003	-7.51
S^2Transformer	Base	0.3911 ± 0.0019	—
	In-sample full	$\mathbf{0.3761} \pm 0.0016$	-3.82
	In-sample matched	0.3840 ± 0.0017	-1.80
	Held-out	0.3786 ± 0.0015	-3.19

Table 23 shows that held-out fitting lowers MSE by 7.79%, 7.51%, and 3.19%, respectively, with improvements in every seed. It also outperforms equal-sized in-sample matched fitting for all three backbones. In-sample full is slightly better but uses 4.07 times as many origins. Thus, adapter gains do not require reusing backbone-training residuals, although temporal distribution differences may also affect the comparison.

C.7 Computational Efficiency

Measurement protocol. We profile seven backbones on France and Europe with seed 2024, batch size 2, and no data-loader workers. Backbone timing measures one training epoch, excluding validation and testing. Adapter timing covers frozen-backbone prediction, training/validation feature accumulation, and ridge fitting with both interaction branches enabled for all six non-MSMV backbones; it excludes prior backbone training and final testing. Comparisons are made within each dataset because batch counts differ. Memory is peak GPU allocation for the complete procedure, including the backbone.

Table 24: Backbone training cost at batch size 2. Time is minutes per training epoch; memory is peak GPU allocation in GiB.

		France		**Europe**	
Paradigm	Backbone	Time (min)	Memory (GiB)	Time (min)	Memory (GiB)
SSSV	xPatch	11.68	0.05	3.02	0.23
	PatchTST	9.95	1.37	1.96	7.56
SSMV	iTransformer	6.48	0.77	2.83	4.18
	DUET	13.81	0.71	3.15	3.56
MSSV	Corrformer	40.67	26.31	30.93	30.03
	CDPNet	141.60	5.68	24.38	2.27
MSMV	HiSTGNN	118.71	2.22	99.80	4.17

Cost of interaction modeling. Table 24 shows that the evaluated MSSV models take substantially longer per epoch than the SSMV models on both datasets, suggesting higher computational costs for station interactions in these implementations. HiSTGNN, the sole native MSMV baseline, is the slowest on Europe (99.80 minutes) and the second slowest on France (118.71 minutes, after CDPNet). Joint modeling can thus incur substantial training cost, although runtime also depends on the backbone architecture.

Table 25: Complete fitting cost of the full dual-branch adapter with a frozen backbone. $T_{\text{fit}}/T_{\text{epoch}}$ compares total adapter fitting time with one training epoch of the same backbone.

Frozen backbone	France			Europe		
	Fit time (min)	Memory (GiB)	Time ratio (%)	Fit time (min)	Memory (GiB)	Time ratio (%)
xPatch	2.77	0.023	23.7	0.64	0.093	21.1
PatchTST	2.33	0.880	23.5	0.80	4.550	40.6
iTransformer	2.93	0.700	45.3	0.62	2.560	22.0
DUET	3.32	0.690	24.1	0.71	2.510	22.6
Corrformer	16.01	28.720	39.4	12.16	21.350	39.3
CDPNet	34.48	5.680	24.3	5.98	2.230	24.5

Adapter fitting cost. As shown in Table 25, complete fitting takes only 21.1–45.3% of one backbone training epoch across all 12 cases, even including frozen-backbone prediction and validation accumulation. This supports the efficiency of the closed-form procedure without iterative backpropagation. Peak memory remains backbone-dependent: it is lower than training in 10 of 12 cases, unchanged at the reported precision for CDPNet on France, and higher for Corrformer on France (28.72 versus 26.31 GiB). Thus, the small adapter parameterization (Appendix B.2) offers consistently low fitting time and usually lower memory requirements, while the total memory footprint also reflects backbone inference and feature processing.

C.8 Sensitivity to Look-Back Window Length

Figure 7 compares seven native backbones on France across look-back windows of 24, 48, 72, and 96 hours, with a fixed 72-hour forecast horizon. Results are averaged across repeated runs and equally over the four target variables.

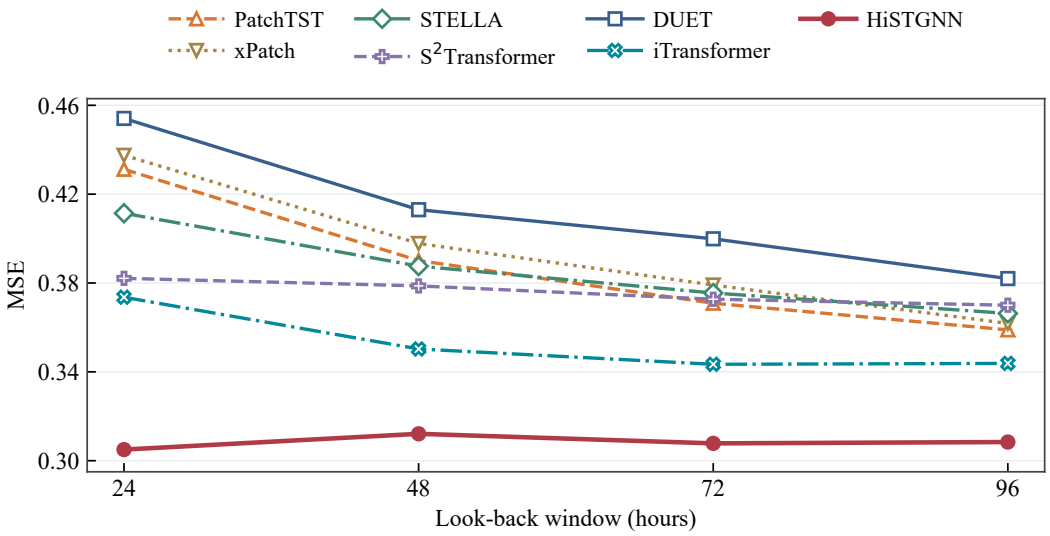

Figure 7: MSE versus look-back window on France with a fixed 72-hour forecast horizon. Each curve reports the mean MSE across repeated runs for a native backbone without an adapter. HiSTGNN maintains the lowest MSE across all four windows.

Longer histories improve most backbones: PatchTST, xPatch, STELLA, S^{2}Transformer, and DUET perform best at 96 hours, whereas iTransformer reaches its minimum at 72 hours. Nevertheless, HiSTGNN consistently achieves the lowest MSE, varying only from 0.3050 to 0.3121. Its advantage therefore persists beyond the default 48-hour setting, supporting the value of joint station–variable modeling across the evaluated look-back windows. The relative ordering of the other paradigms varies with window length, indicating that sensitivity to historical context remains model-dependent.

C.9 Sensitivity to KNN Neighborhood Size

We vary the number of neighbors $K \in \{2, 4, 8, 16\}$ in the station-interaction adapter for iTransformer, TQNet, DUET, and Timer-XL on France and Global. The backbone remains frozen, and the ridge coefficient is selected on the validation set for each configuration. Figure 8 reports test MSE averaged over three runs and the four target variables.

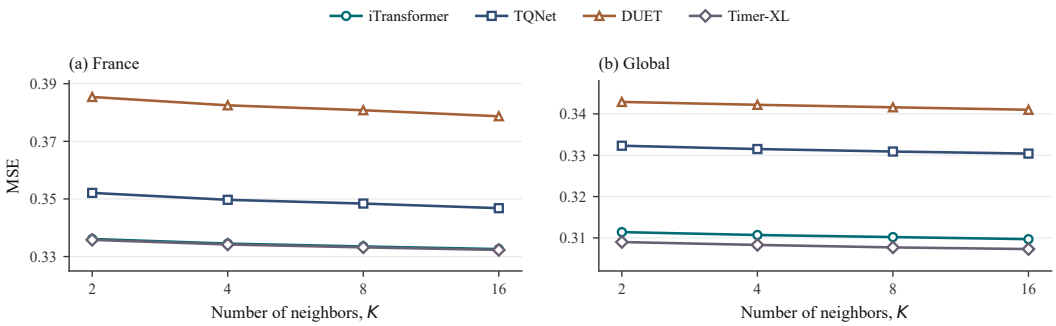

Figure 8: Sensitivity of the station-interaction adapter to the KNN neighborhood size K. Each curve shows the three-run mean MSE averaged over four variables; lower is better.

MSE decreases as K increases for all four backbones on both datasets, with smaller changes from $K = 8$ to 16 than from $K = 2$ to 8 in most cases. Every tested configuration improves upon its corresponding frozen backbone. This indicates that the benefit of station interaction is stable across the examined neighborhood sizes. Considering both efficiency and performance, we use $K = 8$ in the main experiments.

C.10 Prediction Visualizations

Figures 9–12 compare native and adapted forecasts for four example windows on Europe, one per target variable. Each figure presents the same station and forecast window for eight backbones: DLinear, PatchTST, xPatch, STELLA, DUET, Corrformer, TQNet, and iTransformer. The plots show 48 hours of observed history followed by a 72-hour forecast in physical units. Gray denotes historical observations, black the future ground truth, red dashed lines the native forecasts, and blue lines the forecasts with the full adapter; the shaded region marks the forecast horizon.

The temperature and humidity examples illustrate improved tracking of a sustained warming trend and a humidity decline that several native forecasts largely miss. In the wind-speed example, adapted forecasts better capture the increase after a prolonged calm period, although short-lived peaks remain smoothed. For pressure, most adapted trajectories follow the slowly varying observations more closely, while STELLA retains substantial fluctuations. These examples illustrate how output correction can improve forecast levels and temporal trends across different backbones, alongside the aggregate results. Together, they provide qualitative evidence for the effectiveness of the adapter across weather variables and backbone architectures.

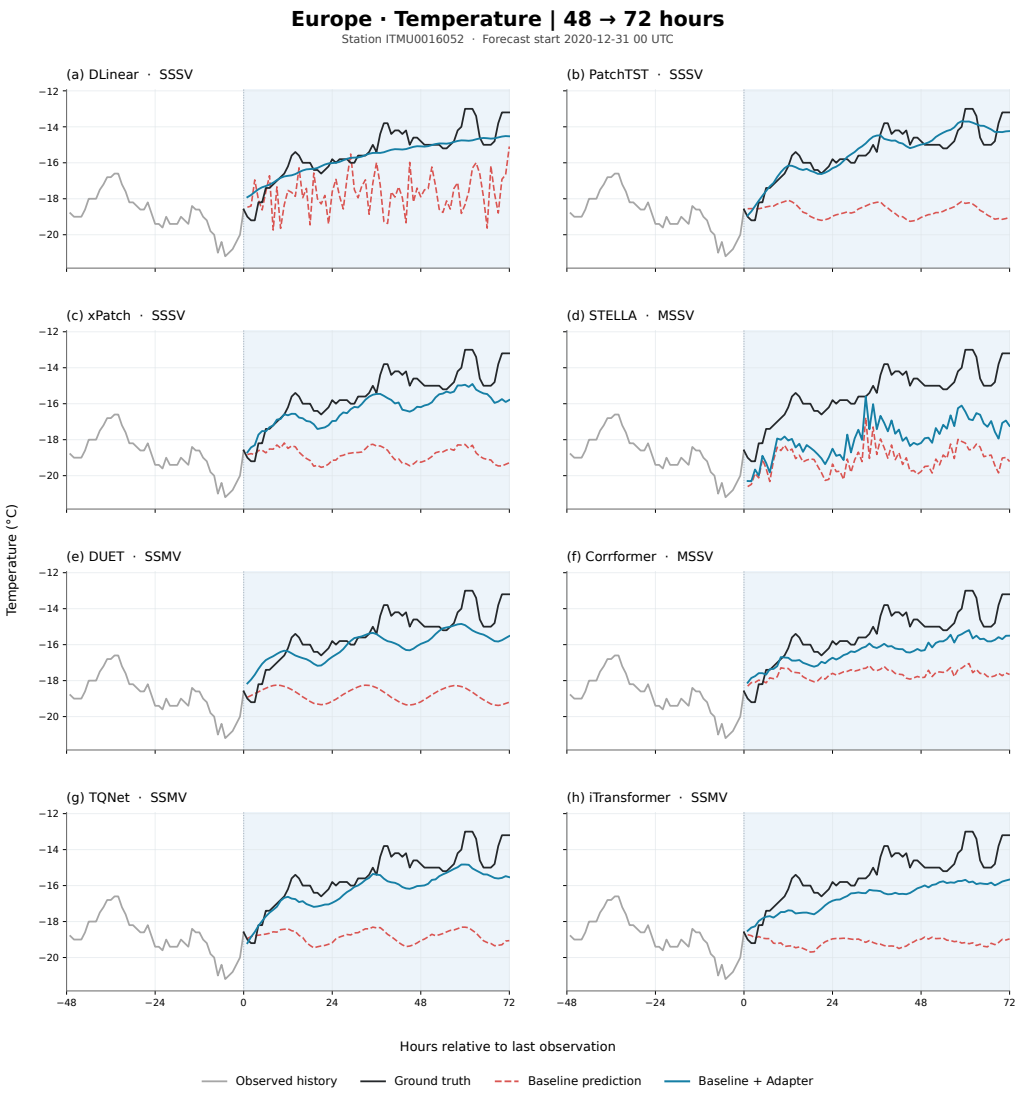

Figure 9: **Air temperature forecasts on Europe.** Station ITMU0016052; forecast start 2020-12-31 00:00 UTC. Each panel compares a native backbone with its full adapter over the same 72-hour window following 48 hours of history. Adapted forecasts better follow the sustained warming trend.

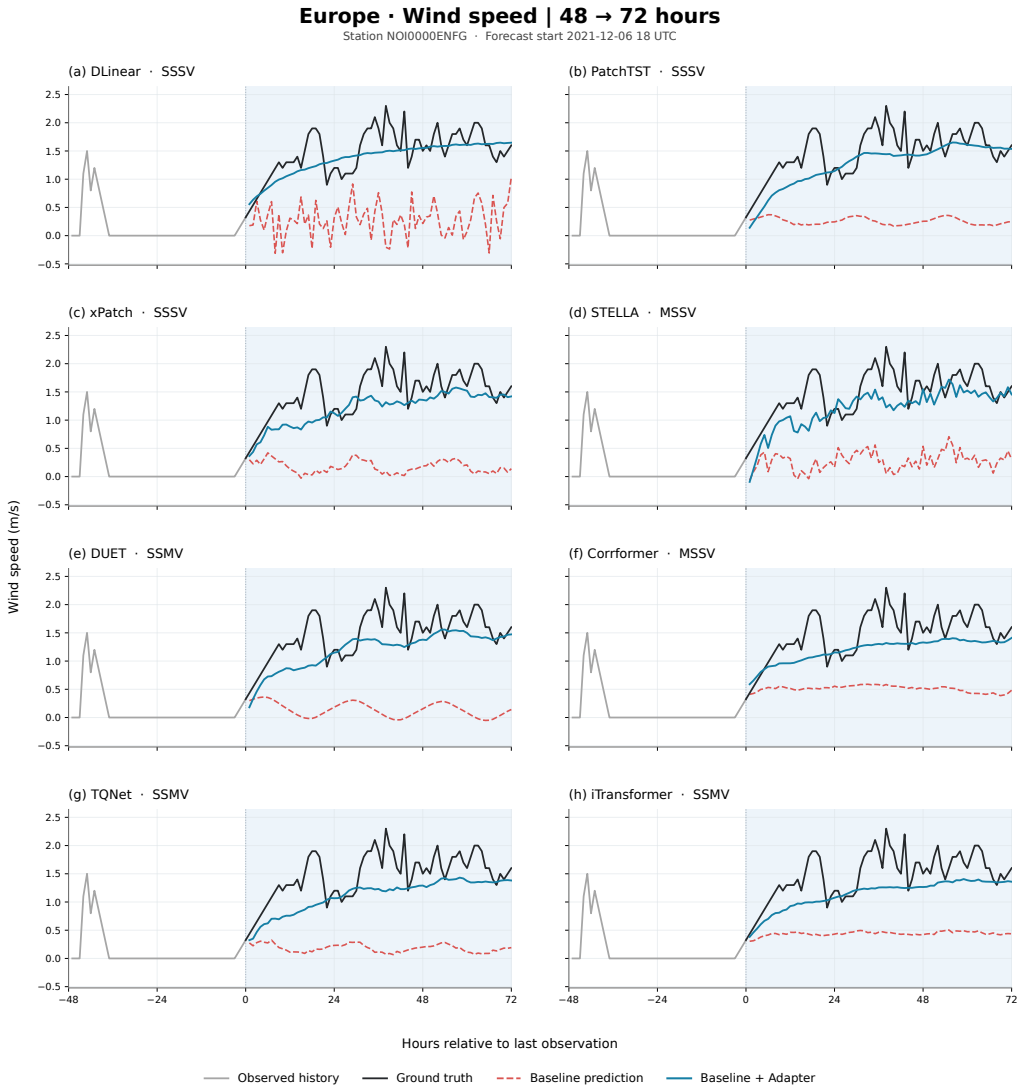

Figure 10: **Wind speed forecasts on Europe.** Station NOI0000ENFG; forecast start 2021-12-06 18:00 UTC. Each panel compares a native backbone with its full adapter over the same 72-hour window following 48 hours of history. Adapted forecasts capture the increase in wind speed, while rapid fluctuations remain difficult to reproduce.

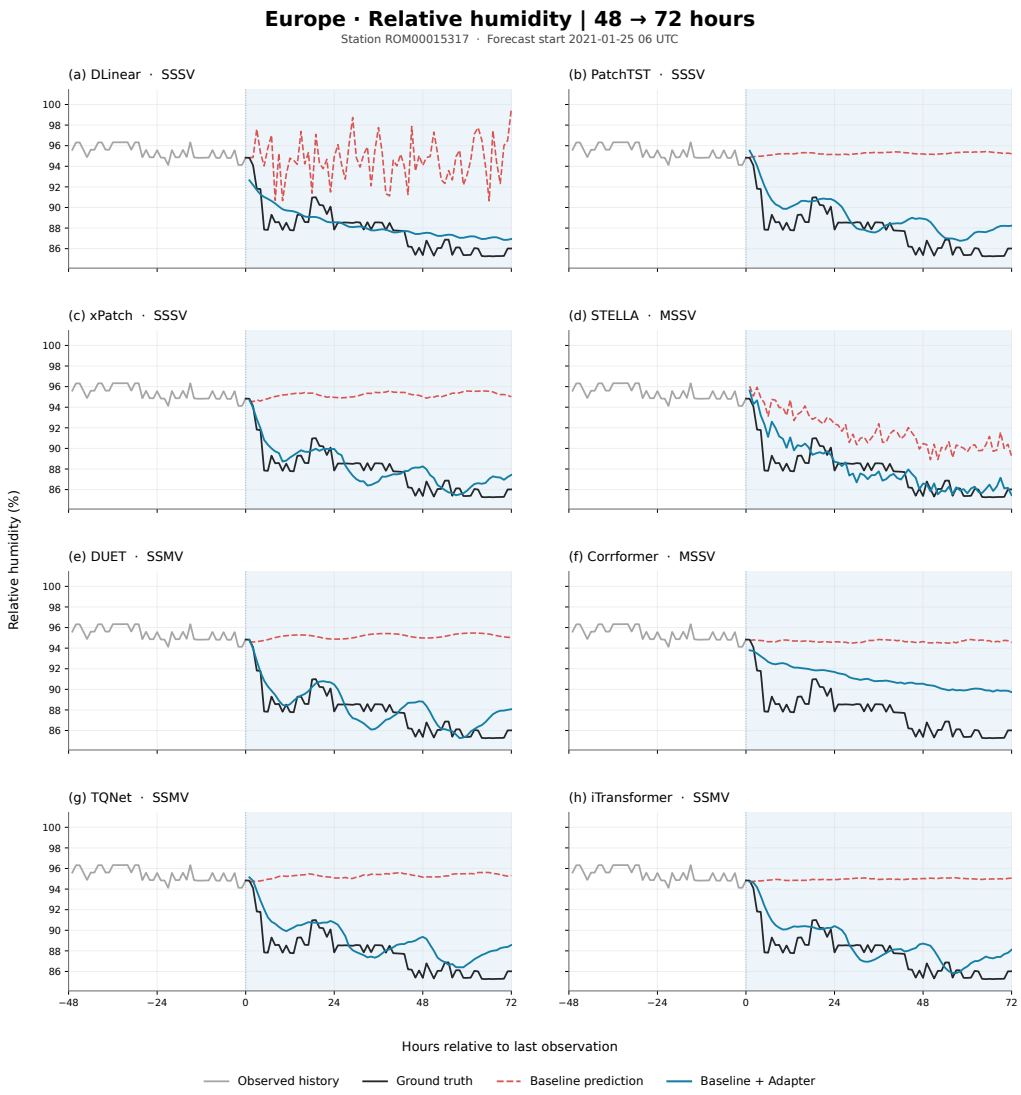

Figure 11: **Relative humidity forecasts on Europe.** Station ROM00015317; forecast start 2021-01-25 06:00 UTC. Each panel compares a native backbone with its full adapter over the same 72-hour window following 48 hours of history. Adapted forecasts better track the humidity decline, with residual differences in timing and magnitude.

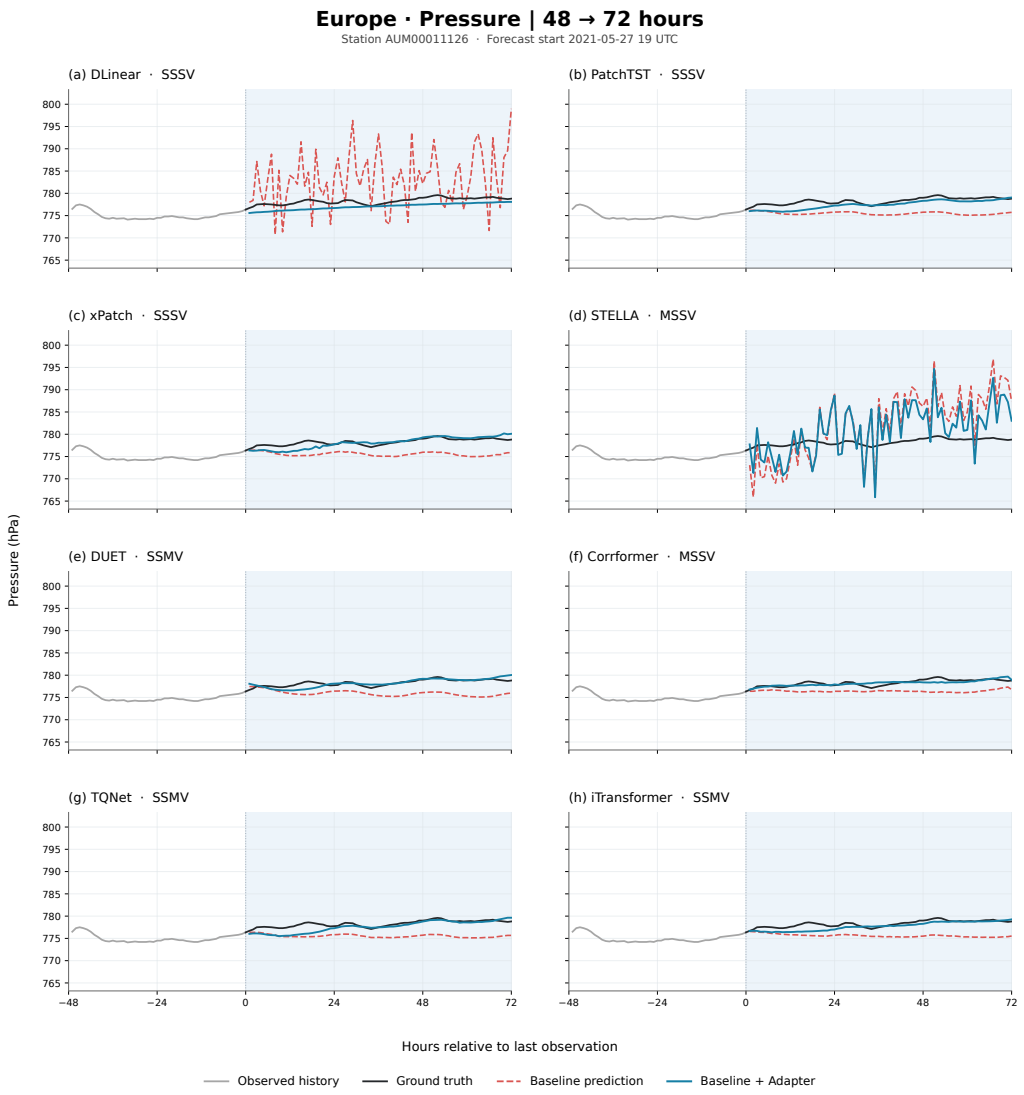

Figure 12: **Station pressure forecasts on Europe.** Station AUM00011126; forecast start 2021-05-27 19:00 UTC. Each panel compares a native backbone with its full adapter over the same 72-hour window following 48 hours of history. Most adapted forecasts follow the observed pressure level and gradual variation more closely; STELLA retains pronounced fluctuations.