TRANSFER-VALID WEATHER-STATION FORECASTING: SEPARATING HELD-OUT COMPUTABILITY FROM SPATIAL COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fixed-station accuracy does not establish transfer to stations excluded from training. We formalize transfer validity for location-known, observation-held-out forecasting as a property of learned prediction under two distinct boundaries. The parameter boundary excludes dependence on identity-indexed trainable state left untrained by the spatial split. Required trainable quantities are instead shared across stations or generated from permitted inputs. The information boundary restricts prediction and routing to data permitted by the split and available at or before the forecast origin. It allows dynamic routes built from the recent history supplied at inference. WeatherScaleNet instantiates these conditions with shared station functions and static-geographic routing. Its completed computational-position graph has diameter at most two when both interaction axes are present, and balanced layouts give spatial attention cost $\mathcal{O}(\widetilde{N}^{3/2}d)$. Five-seed experiments on global weather stations and physical rain gauges evaluate fixed-station accuracy and random-station and contiguous-region holdout. Fixed-station branch controls favor retaining both interaction axes. Coordinate and elevation representations change rank across settings, feature concatenation can outperform scalar aggregation, and partition granularity has limited influence over the tested range. These findings distinguish learned held-out computation, permitted information use, communication structure, and predictive effectiveness without tying transfer validity to one encoder or routing mechanism.

1 INTRODUCTION

Fixed-station forecasting evaluates the same station identities during training and testing. It measures temporal generalization at recurring nodes, but does not establish performance at a station excluded from optimization. A station-indexed embedding can be highly predictive when identities recur. At a held-out station, its row remains at initialization. That row can still produce a forecast, but supplies no learned state for the excluded identity.

We distinguish two aspects of held-out computation. The *parameter boundary* asks whether prediction relies on shared learned quantities or instead requires identity-specific trainable state left untrained by the split. The *information boundary* asks whether prediction and routing use only data permitted by the split and forecast origin. The first concerns learned transfer; the second concerns information leakage. We use *transfer validity* for computations satisfying both, and evaluate their predictive effectiveness separately.

Location-known, observation-held-out forecasting has a specific information schedule. Coordinates and elevation remain available as static metadata, while held-out histories and targets do not contribute to parameter optimization. At inference, each held-out station supplies a recent observation window and missingness mask. A route constructed from this permitted history can respect the protocol. A route estimated during training from excluded observations cannot.

WeatherScaleNet is one instantiation. Shared history, coordinate, and elevation functions construct station tokens. Coordinate partitions and geographic-distance completion define a static route.

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.

Within-partition and cross-partition attention then exchange information along two computational axes. This partitioned two-axis attention pattern is related to prior structured spatial models (Fang et al., 2025) and enables us to study communication structure separately from the parameter and information boundaries.

Our contributions are threefold.

1. We formalize transfer validity for location-known, observation-held-out forecasting, separating learned held-out computation from protocol-respecting information use and predictive accuracy.
2. We verify these conditions for WeatherScaleNet and characterize reachability and spatial attention cost on its completed computational-position graph.
3. We find that fixed-station branch controls favor both communication axes, whereas coordinate and elevation representations change rank across settings and aggregation choices trade accuracy for model size. Random-station and contiguous-region holdouts evaluate the resulting forecasts separately from these communication controls.

2 Related Work

2.1 Station forecasting and held-out locations

Station-wise statistical and neural predictors model temporal variation without exchanging information among locations. Graph and attention models address this limitation through geographic or learned relations. HiSTGNN combines station-level and regional graph representations (Ma et al., 2023). Corrformer models spatial cross-correlation and temporal autocorrelation in a hierarchical weather forecasting system (Wu et al., 2023). These methods establish the value of structured spatial exchange for recurring nodes.

2.2 Inductive prediction at held-out nodes

Inductive graph learning generates representations for unseen nodes through shared functions and node features rather than a separately trained embedding for every node (Hamilton et al., 2017). Inductive spatiotemporal kriging likewise studies prediction at unsampled sensors and transfer across graph structures (Wu et al., 2021b). Observation-held-out forecasting has a distinct information schedule. A target station is absent during optimization but can provide a recent history at forecast time. Representation inductiveness and route admissibility are therefore separate questions. A shared encoder can satisfy the parameter boundary while a graph estimated from held-out training observations violates the information boundary. Conversely, a dynamic graph computed from the permitted forecast-time history can satisfy both boundaries. Transfer validity audits these two failure modes separately.

2.3 Coordinate and terrain representations

Coordinates can be supplied directly, expanded through Fourier features, or mapped through a pretrained location model. GeoCLIP learns a worldwide location representation from image-location alignment (Cepeda et al., 2023). WeatherScaleNet uses only its location branch and receives no image at inference. Raw coordinates, Fourier features, a random fixed vector, station identity, and no coordinate representation are evaluated as controls. Elevation is represented by empirical-quantile bins or a continuous network. Shared geospatial functions can therefore be implemented in several ways.

2.4 Structured spatial communication

Full attention connects every station pair at quadratic spatial cost (Vaswani et al., 2017). Earthformer reduces spatiotemporal attention cost through structured cuboids on gridded data (Gao et al., 2022). Other forecasting systems use partitions, subgraphs, and hierarchical exchange to balance regional and network-wide interaction.

PatchSTG is the closest operation-level precedent. It partitions irregular points with a spatial tree, completes unequal patches, applies attention within patches, and applies attention across equal in-

dices of different patches (Fang et al., 2025). WeatherScaleNet uses this established partitioned attention pattern. S^{2}Transformer exchanges information within and between station subgraphs while changing subgraph scale (Chen et al., 2025). SIGMAformer constructs dynamic temporal and spatial correlations from Gaussian mixture patterns (Kim and Suk, 2026).

These works provide partitioned, hierarchical, and dynamic communication mechanisms. Our analysis separates their learned parameterization from the information used to construct their routes. A shared representation and a protocol-respecting route address different conditions; neither alone determines forecast accuracy.

3 Problem Setting and Transfer Validity

3.1 Fixed-station and held-out-station forecasting

Let $\mathcal{V} = \{1, \ldots, N\}$ be a station set. Station i has coordinates $\mathbf{g}_i$, elevation h_i, observations $\mathbf{y}_{i,t}$, and a binary observation mask $\mathbf{m}_{i,t}$. At forecast origin τ, a model receives a look-back window of length L_{in}:

$$\mathcal{H}_\tau = \{\mathbf{y}_{i,\tau-L_{\text{in}}+1:\tau}, \mathbf{m}_{i,\tau-L_{\text{in}}+1:\tau}, \mathbf{g}_i, h_i\}_{i\in\mathcal{V}_{\text{in}}}. \tag{1}$$

It predicts L_{out} steps for $\mathcal{V}_{\text{out}}$:

$$\widehat{\mathbf{Y}}_{\tau+1:\tau+L_{\text{out}}} = F_\Theta\left(\mathcal{H}_\tau, \mathcal{V}_{\text{out}}\right). \tag{2}$$

In fixed-station evaluation, the same stations appear during training and testing, and time windows are separated chronologically. In spatial transfer, the station set is divided into training stations $\mathcal{V}_{\text{train}}$ and held-out stations $\mathcal{V}_{\text{held}}$. Histories and targets from $\mathcal{V}_{\text{held}}$ do not contribute to parameter optimization. Coordinates and elevation remain available as static metadata. At validation and test time, each held-out station supplies its recent look-back observations and mask. Metrics are computed only on $\mathcal{V}_{\text{held}}$.

For a target set $\mathcal{V}_*$, the masked objective is

$$\mathcal{L}(\Theta) = \frac{\sum_\tau \sum_{i\in\mathcal{V}_*} \sum_{k=1}^{L_{\text{out}}} m_{i,\tau+k}\, \ell\left(\widehat{y}_{i,\tau+k}, y_{i,\tau+k}\right)}{\sum_\tau \sum_{i\in\mathcal{V}_*} \sum_{k=1}^{L_{\text{out}}} m_{i,\tau+k}}. \tag{3}$$

Missing values can be filled for tensor storage, but they do not contribute to fitting or evaluation.

3.2 Transfer validity

Definition 1. Transfer validity. *Under location-known, observation-held-out forecasting, a forecaster is transfer-valid if, for every held-out station, the following conditions hold.*

i. ***Held-out computability.*** *Every trainable quantity required for the prediction is shared across stations or generated by a shared learned function of inputs permitted for that station. The prediction does not require an identity-indexed trainable parameter whose held-out value was left untrained by the spatial split. This is the parameter boundary.*
ii. ***Protocol-respecting routing.*** *Prediction inputs, spatial routes, and auxiliary statistics use only model quantities and information permitted by the evaluation protocol at or before the forecast origin. Parameter fitting and statistic estimation exclude observations withheld during training, and forecast-time computation excludes future held-out targets. This is the information boundary.*

Transfer validity characterizes whether a held-out prediction is supported by learned computation and protocol-permitted information; it is independent of predictive accuracy.

3.3 Claim and evidence hierarchy

Table 1 separates four claims that require different evidence.

Table 1: Four evidence levels in held-out station forecasting.

Level	Question	Evidence	Supported claim
Transfer validity	Is the held-out computation transfer-valid?	Definition 1 and Proposition 1	Learned held-out computation with protocol-respecting information
Communication structure	What can the completed graph communicate, and at what cost?	Propositions 2 and 3	Bounded-depth reachability and subquadratic spatial attention under balanced layouts
Communication effectiveness	Does the mechanism improve forecasting?	Fixed-station branch controls	The recorded branch ordering favors retaining both interaction axes
Implementation choice	Which realization works best?	Coordinate, elevation, aggregation, partition, robustness, and efficiency controls	No tested realization is uniformly best across all settings

Table 2: Transfer-validity audit examples. All unspecified inputs respect the protocol. Recent history denotes the permitted look-back window.

Forecaster construction	Learned held-out computation	Protocol-respecting information	Transfer-valid
Shared backbone requiring a trainable station-identity row left untrained by the split	No	Yes	No
Shared encoder with a graph estimated from the held-out training sequence	Yes	No	No
Shared encoder with a static coordinate graph	Yes	Yes	Yes
Shared encoder with a dynamic graph built from permitted recent history	Yes	Yes	Yes

4 A Transfer-Valid Instantiation: WeatherScaleNet

WeatherScaleNet satisfies Definition 1 through shared station functions and a static-geographic route. Its route uses coordinates and geographic distance, a subset of the information permitted by the definition.

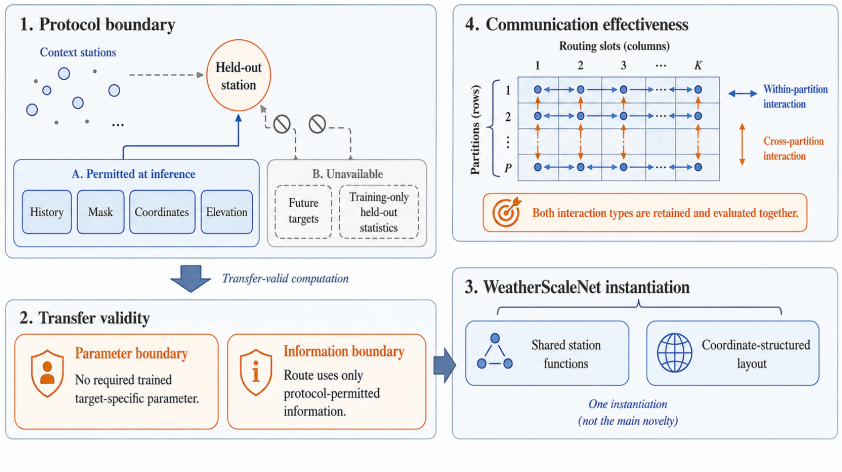

Figure 1: Transfer validity and communication design. The parameter boundary concerns learned held-out computation; the information boundary concerns permitted data use. WeatherScaleNet satisfies both through shared station functions and a static coordinate-structured layout. Graph analysis characterizes the two interaction axes, and forecasting controls evaluate their effectiveness. Routing slots are computational indices, not predefined physical correspondences.

4.1 SHARED STATION FUNCTIONS

For station i and forecast origin τ, the station token is

$$\mathbf{x}_{i,\tau} = \left[f_{\text{history}}\left(\mathbf{y}_{i,\tau-L+1:\tau}, \mathbf{m}_{i,\tau-L+1:\tau}\right) \parallel f_{\text{coordinate}}\left(\mathbf{g}_i\right) \parallel f_{\text{elevation}}\left(h_i\right) \right]. \tag{4}$$

All functions are shared across stations. The history encoder uses non-overlapping temporal patches. The default coordinate function is the frozen GeoCLIP location branch followed by a learned projection (Cepeda et al., 2023). The default elevation function uses empirical-quantile bins. Raw coordinates, Fourier features, a random fixed vector, station identity, no coordinate representation, and a continuous elevation network are evaluated as alternatives. No raster image, satellite image, or required station-indexed trainable parameter enters the default computation.

The global and MeteoNet configuration uses 64 history dimensions, 32 coordinate dimensions, and 32 elevation dimensions. The complete token dimension is 128. A temporal patch length of 24 maps a 48-hour history to two tokens. The daily precipitation configuration uses a temporal patch length of 15 and maps a 30-day history to two tokens.

4.2 COORDINATE-STRUCTURED LAYOUT

A balanced two-dimensional k-d tree recursively bisects stations at coordinate medians and alternates latitude and longitude. The leaves are spatially compact and similar in size. Consecutive leaves are combined through a merge factor to form P effective partitions. Each partition is extended to S computational positions by copying context stations when needed. Fixed-station experiments estimate copying affinity from training observations only. Spatial-holdout experiments use haversine distance only. Held-out observations and forecast targets do not enter route construction.

The routing-slot order follows the implementation. Recursion visits the lower-coordinate child before the upper-coordinate child. Station indices within a leaf remain in ascending input order. Effective partitions concatenate consecutive leaves in recursive order. The route is therefore coordinate-structured but not purely coordinate-defined. It is deterministic for a fixed station array and is not invariant to a permutation of the input indices.

After context completion, one temporal slice has layout

$$\mathbf{X} \in \mathbb{R}^{P \times S \times d}. \tag{5}$$

Copied occurrences participate as separate computational positions. The inverse index map later discards their repeated outputs and restores the original station order.

4.3 TWO-AXIS SPATIAL COMMUNICATION

Within each partition, self-attention is applied over its S routing slots:

$$\mathbf{X}^{\text{within}}_{p,:,:} = \text{Block}_{\text{within}}\left(\mathbf{X}_{p,:,:}\right). \tag{6}$$

The corresponding computational-position edges are

$$E_{\text{within}} = \left\{ \left((p,s),(p,s')\right) : s \neq s' \right\}. \tag{7}$$

The partition and slot axes are then transposed. Self-attention is applied across P partitions for each routing slot:

$$\mathbf{X}^{\text{cross}}_{:,s,:} = \text{Block}_{\text{cross}}\left(\mathbf{X}_{:,s,:}\right). \tag{8}$$

The corresponding edges are

$$E_{\text{cross}} = \left\{ \left((p,s),(p',s)\right) : p \neq p' \right\}. \tag{9}$$

Equal slots define a sparse computational route. They do not assert geographic proximity or a known climatic correspondence.

4.4 AGGREGATION AND FORECAST DECODER

The default implementation aggregates the two paths with normalized learned scalars:

$$(\alpha_{\text{within}}, \alpha_{\text{cross}}) = \text{softmax}\left(\omega_{\text{within}}, \omega_{\text{cross}}\right), \tag{10}$$

$$\mathbf{Z} = \alpha_{\text{within}}\mathbf{X}^{\text{within}} + \alpha_{\text{cross}}\mathbf{X}^{\text{cross}}. \tag{11}$$

Equal initialization starts both weights at 0.5. The aggregation keeps the representation dimension fixed and makes path-weight collapse observable. It is not a defining condition of transfer validity. Feature concatenation followed by projection is a higher-capacity alternative. A shared convolutional decoder maps the recovered station representations to all forecast lead times.

The reported network uses five spatial layers, one attention head, query-key-value bias, attention dropout 0.1, projection dropout 0.1, affine-free layer normalization, and a feed-forward expansion ratio of one.

5 STRUCTURAL ANALYSIS

The analysis verifies the two boundaries and characterizes reachability and cost on the completed computational layout.

Proposition 1. Transfer validity of WeatherScaleNet. *WeatherScaleNet is transfer-valid under the stated location-known, observation-held-out protocol.*

Proof. **Learned held-out computation.** The history, coordinate, and elevation functions, spatial blocks, and decoder share their trainable weights across stations. They generate the held-out representation from permitted inputs, without requiring an identity-indexed trainable row left untrained by the split. The parameter boundary is satisfied.

Protocol-respecting information. Parameter optimization excludes held-out histories and targets. Forecast-time inputs contain only the permitted recent histories, masks, and static metadata. The spatial-holdout partition uses coordinates, and context completion uses geographic distance. No route statistic is estimated from excluded training observations or future targets. The information boundary is satisfied, completing the verification of Definition 1. □

Let

$$\widetilde{\mathcal{V}} = \{(p, s) : 1 \le p \le P,\ 1 \le s \le S\} \tag{12}$$

be the computational positions in the completed layout. Two copied occurrences of the same station are distinct vertices in this graph. Only the recovery map identifies which outputs correspond to original stations.

Proposition 2. Reachability of the completed layout. *Consider the computational-position graph on $\widetilde{\mathcal{V}}$. With only E_{within}, the graph has P disconnected components. With only E_{cross}, it has S disconnected components. With both edge sets, its diameter is at most two.*

Proof. Within-partition edges preserve the partition index, so each partition forms a separate component. Cross-partition edges preserve the routing-slot index, so each slot forms a separate component. Consider positions (p, s) and (p', s'). They are adjacent when $p = p'$ or $s = s'$. Otherwise,

$$(p, s) \to (p, s') \to (p', s') \tag{13}$$

is a path of length two. The two paths are computed in parallel within one spatial layer, so arbitrary computational positions can influence one another after two stacked spatial layers. □

Proposition 3. Spatial attention cost. *The within-partition path has cost $\mathcal{O}(PS^2d)$ and the cross-partition path has cost $\mathcal{O}(SP^2d)$. The total spatial attention cost is*

$$\mathcal{O}\left((PS^2 + SP^2)d\right). \tag{14}$$

Let $\widetilde{N} = PS$ denote the number of computational positions in the completed layout. If $P, S = \Theta(\sqrt{\widetilde{N}})$, the total cost is $\mathcal{O}(\widetilde{N}^{3/2}d)$, below the $\mathcal{O}(\widetilde{N}^2d)$ cost of full attention over the completed layout. If context completion adds at most a constant-factor number of positions relative to the original station count, the same asymptotic scaling holds in the number of stations.

6 EXPERIMENTAL SETUP

Global Wind and Global Temperature each contain 3,850 physical stations. MeteoNet NW and SE contain 231 and 451 physical rain gauges. These tasks use 48 input hours and 24 forecast hours. Four regular one-degree precipitation meshes provide a separate extension with 30 input and 30 forecast days. Appendix B specifies data processing, splits, optimization, metrics, and baselines.

Fixed-station evaluation uses chronological splits. Random-station holdout excludes 20 percent of stations, while contiguous-region holdout selects a coordinate block along a seed-dependent projection axis. Held-out histories and labels are excluded from optimization; recent histories enter at inference. Metrics are computed on held-out targets, and WeatherScaleNet uses geographic routing in these tests.

Controlled experiments use five seeds and report a two-sided 95 percent t interval. Paired analyses use matched seeds. Conditional temporal intervals use 5,000 bootstrap repetitions over complete non-overlapping blocks of 168 forecast origins within each seed. The controlled comparison emphasizes mean absolute error, while the broader benchmark retains mean absolute error, mean squared error, and root mean squared error. Rainfall-event scores, robustness diagnostics, and efficiency are reported in the appendices.

7 MAIN RESULTS

7.1 RQ1: CAN A TRANSFER-VALID INSTANTIATION RETAIN STRONG FIXED-STATION ACCURACY?

We first test whether a transfer-valid instantiation can retain strong fixed-station accuracy. In the matched five-seed comparison in Table 3, WeatherScaleNet obtains lower mean absolute error than the rerun STELLA and TimeMixer baselines on both targets.

Table 3: Controlled fixed-station forecasting. Entries are five-seed mean absolute error with a 95 percent t interval. Lower is better.

Dataset	WeatherScaleNet	STELLA	TimeMixer	ST-Balance
Global Wind	1.2066 ± 0.0023	1.3174 ± 0.0002	1.3271 ± 0.0041	1.258 ± 0.0012
Global Temperature	1.5944 ± 0.0066	1.9310 ± 0.0018	1.9005 ± 0.0021	1.831 ± 0.0031

For Global Wind, the paired mean absolute error differences are 0.1108 for STELLA and 0.1206 for TimeMixer. For Global Temperature, they are 0.3366 and 0.3061. The paired seed intervals and conditional block intervals exclude zero in every comparison.

Table 4 compares simple, statistical, numerical, graph-based, temporal, and meteorological methods under the single-run benchmark protocol. WeatherScaleNet has the lowest MAE, MSE, and RMSE on both targets. Among the alternatives, Corrformer has the lowest wind MAE and temperature squared error, TimeMixer has the lowest wind squared error, and CATS has the lowest temperature MAE.

The single-run benchmark describes comparison breadth; the five-seed runs support uncertainty and paired inference. Their values are not pooled. The benchmark includes GFS and ERA5 as external references rather than model inputs, with reanalysis distinguished from forecasts.

Appendix C gives complete controlled intervals and spatial diagnostics. Appendix Table 14 reports all four regular-mesh precipitation benchmarks separately from physical-station transfer.

7.2 RQ2: CAN A TRANSFER-VALID MODEL FORECAST STATIONS EXCLUDED FROM OPTIMIZATION?

Table 5 reports the default coordinate representation under random-station and contiguous-region holdout. Proposition 1 verifies the learned-computation and information conditions. The experiment measures forecasting feasibility and error after target identities are excluded from optimization and geographic separation increases.

Table 4: Broader fixed-station benchmark under the established single-run protocol. Values are reported in physical units. Lower is better. The controlled five-seed comparison in Table 3 provides uncertainty estimates.

Method	Global Wind			Global Temperature		
	MAE	MSE	RMSE	MAE	MSE	RMSE
Simple baselines						
Repeat last time	1.611	5.745	2.397	3.137	20.850	4.566
Repeat last period	1.746	6.955	2.637	2.471	13.306	3.648
Statistical methods						
Holt-Winters	1.664	5.912	2.431	2.262	13.241	3.639
GBDT	1.417	4.101	2.025	2.214	9.706	3.115
Prophet	2.382	9.691	3.113	2.946	11.626	3.410
Numerical products						
GFS prediction at 0.25°	2.340	9.993	3.161	2.287	14.933	3.864
GFS reanalysis at 0.5°	1.940	6.767	2.601	2.254	14.809	3.848
ERA5 prediction at 0.5°	1.847	6.793	2.606	3.270	28.069	5.297
ERA5 reanalysis at 0.25°	1.587	4.999	2.236	1.900	13.448	3.667
Graph-based models						
StemGNN (Cao et al., 2020, NeurIPS)	1.389	4.066	2.016	2.746	13.926	3.730
Graph WaveNet (Wu et al., 2019, IJCAI)	1.342	3.912	1.978	2.438	11.724	3.424
ST-Balance (Chen et al., 2026, Neural Networks)	1.289	3.551	1.884	1.848	7.624	2.761
Temporal forecasting models						
FEDformer (Zhou et al., 2022, ICML)	1.503	4.750	2.179	2.405	11.054	3.325
CATS (Kim et al., 2024, NeurIPS)	1.339	3.923	1.981	1.878	7.862	2.804
TimeMixer (Wang et al., 2024, ICLR)	1.329	3.801	1.950	1.940	7.963	2.822
Meteorological forecasting models						
Corrformer (Wu et al., 2023, Nature Machine Intelligence)	1.304	3.889	1.972	1.888	7.709	2.777
MRIformer (Yu et al., 2024, Information Sciences)	1.318	3.906	1.976	1.999	9.516	3.085
STELLA (Fu et al., 2026, NeurIPS)	1.311	3.810	1.952	2.017	9.373	3.061
WeatherScaleNet	**1.190**	**3.151**	**1.775**	**1.510**	**5.249**	**2.291**

Table 5: Location-known, observation-held-out forecasting with the default coordinate representation. Mean absolute error is the five-seed mean with a 95 percent t interval. Distance is the mean distance from a held-out station to its nearest training station.

Target	Random MAE	Block MAE	Random distance, km	Block distance, km
Global Wind	1.3010 ± 0.0151	1.3064 ± 0.0639	96.8	2801.7
Global Temperature	1.9291 ± 0.0428	2.1183 ± 0.2719	96.8	2801.7
MeteoNet NW precipitation	0.1174 ± 0.0036	0.1168 ± 0.0068	19.2	72.1
MeteoNet SE precipitation	0.1639 ± 0.0068	0.1685 ± 0.0192	13.0	89.2

The contiguous split greatly increases nearest-training distance, especially in the global network. Larger distance does not guarantee larger average error because each seed selects a different edge region with a different target distribution. The wider global temperature interval under block holdout reflects this regional variation. MeteoNet results show feasibility with native observation gaps, while stronger rain thresholds remain difficult. Appendix D reports station maps, complete MeteoNet event scores, and station-level error maps.

7.3 RQ3: Are both communication axes useful in fixed-station controls?

Proposition 2 motivates testing whether the two communication axes are empirically complementary. On both global targets, retaining both axes gives lower five-seed mean absolute error than either the within-partition-only or cross-partition-only variant. Paired t analyses support this ordering. Appendix E gives the statistical details.

7.4 RQ4: Which choices are defining conditions, and which are replaceable?

Table 6 isolates coordinate and identity representations in global spatial transfer. No representation dominates all targets and splits. The pretrained representation has the lowest mean wind error under both holdouts, with a small margin over raw coordinates. Raw coordinates have the lowest temperature error under random holdout. Omitting a position representation has the lowest temperature error under block holdout.

Table 6: Coordinate and identity representations in global spatial transfer. Twenty percent of stations are excluded from optimization. Entries are five-seed mean absolute error with a 95 percent t interval.

	Global Wind		Global Temperature	
Representation	Random	Block	Random	Block
No position representation	1.3071 ± 0.0250	1.3182 ± 0.0609	1.8975 ± 0.0533	2.0141 ± 0.2307
Raw latitude and longitude network	1.3045 ± 0.0181	1.3091 ± 0.0943	1.8627 ± 0.0652	2.0395 ± 0.2174
Random Fourier features	1.3077 ± 0.0240	1.3132 ± 0.0579	1.9466 ± 0.0558	2.1467 ± 0.2189
Preallocated station identity	1.3784 ± 0.0212	1.3865 ± 0.0886	1.9716 ± 0.0240	2.1004 ± 0.2360
Pretrained coordinate representation	1.3010 ± 0.0151	1.3064 ± 0.0639	1.9291 ± 0.0428	2.1183 ± 0.2719

Held-out identity rows remain at initialization. Their scores are valid benchmark measurements under the information boundary, but the required untrained rows fail the parameter boundary. Shared coordinate functions provide learned representations for held-out locations without guaranteeing that a particular encoder will be most accurate.

Elevation bins are slightly stronger for wind, while a continuous network is slightly stronger for temperature. Concatenation gives lower global error than scalar aggregation, which uses fewer parameters and preserves the representation dimension. Error changes little across the tested partition counts. Appendix E gives the grouped and factorial controls; the remaining appendices report robustness, precipitation, and efficiency.

8 Discussion

Transfer validity separates two requirements for learned forecasting at unseen stations: the model must construct a learned representation for the held-out station, and its computation must respect the forecast-time information boundary. This separation is especially relevant in observation networks, where a station can be absent during optimization yet return with recent measurements at inference.

The experiments show that satisfying these conditions does not select a unique spatial representation or communication implementation. Shared coordinate functions remain applicable to held-out locations, but the most accurate encoding varies across targets and spatial splits. In the fixed-station controls, both communication axes are useful together, while concatenation offers an accuracy-oriented alternative to scalar aggregation. The graph analysis explains the reachability and attention cost of the two-axis layout.

Routing slots depend partly on input indices within k-d tree leaves, so routes are not permutation invariant. Fixed partitions can miss changing atmospheric relations. Native missingness shifts and rare heavy precipitation remain difficult for the deterministic forecasts, and two-path interaction adds training cost.

9 Conclusion

Transfer validity separates learned held-out computation from protocol-respecting information use. A held-out prediction must not require identity-indexed trainable state left untrained by the spatial split. Shared or input-generated quantities satisfy this parameter condition, while the information boundary independently restricts prediction and routing to permitted data.

WeatherScaleNet provides one realization through shared station functions and static-geographic routing. Its completed two-axis graph has bounded-depth reachability and subquadratic attention cost under balanced layouts. Fixed-station controls favor both axes within this communication family; held-out experiments evaluate forecasting at excluded identities. Encoder, elevation, aggregation, and partition choices remain dataset-dependent rather than defining conditions of transfer validity.

Reproducibility Statement

The experiment archive records data processing, chronological and spatial splits, random seeds, configurations, checkpoints, run-level metrics, paired tests, forecast-origin errors, environment manifests, partition index maps, and node-order recovery maps. An anonymized repository and result archive will accompany the submission.

AI Use Statement

Generative AI tools were used for language editing only. The authors reviewed all content and take full responsibility for the manuscript.

References

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

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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. *arXiv preprint arXiv:2509.19648*, 2025.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

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*, 2026.

Zhangyang Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yung B. Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for Earth system forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pages 25390–25403, 2022.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dongbin Kim, Jinseong Park, Jaewook Lee, and Hoki Kim. Are self-attentions effective for time series forecasting? In *Advances in Neural Information Processing Systems*, volume 37, pages 114180–114209, 2024.

Do-Yeon Kim and Heung-Il Suk. SIGMAformer: A spatiotemporal Gaussian mixture correlation transformer for global weather forecasting. *npj Climate and Atmospheric Science*, 9:113, 2026.

Météo-France. MeteoNet: An open reference meteorological dataset. Official dataset documentation, 2020.

Ming Ma, Peng Xie, Fei Teng, Bin Wang, Shihao Ji, Jian Zhang, and Tao Li. HiSTGNN: Hierarchical spatio-temporal graph neural network for weather forecasting. *Information Sciences*, 648:119580, 2023.

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

Adam Smith, Neal Lott, and Russell Vose. The Integrated Surface Database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92:704–708, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1907–1913, 2019.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021.

Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4478–4485, 2021.

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.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3634–3640, 2018.

Chengqing Yu, Guangxi Yan, Chengming Yu, Xinwei Liu, and Xiwei Mi. MRIformer: A multi-resolution interactive transformer for wind speed multi-step prediction. *Information Sciences*, 661:120150, 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the International Conference on Machine Learning*, volume 162, pages 27268–27286, 2022.

Jing Chen and Shixiang Pan and Yujie Fan and Haocheng Ye and Haitao Xu and Wenqiang Xu. Dimensional balance improves large scale spatiotemporal prediction performance. *Neural Networks*, 205:109434, 2026.

A Implementation Details

A.1 Shared representations

The history encoder is a shared two-dimensional convolution with a spatial kernel of one station. Its temporal kernel and stride equal the temporal patch length. Global and MeteoNet tasks use a temporal patch length of 24. Daily gridded precipitation uses a temporal patch length of 15.

The default coordinate representation is produced by the GeoCLIP location branch and a learned linear projection. GeoCLIP is evaluated as a frozen feature generator. The distributed package version is 1.2.0 and the checkpoint checksum prefix is `94b80ae3af89`. Alternative controls replace the default with raw latitude and longitude, standalone Random Fourier features, a random fixed 512-dimensional vector, no position representation, or a preallocated station-identity table. The coordinate controls are mutually exclusive.

Elevation is divided into 50 empirical-quantile bins. Duplicate bin edges are removed. If the observed elevations do not define all bins, equally spaced edges over the observed elevation range are used. Missing elevation is replaced by the network median before binning. A continuous elevation network is evaluated as an alternative.

A.2 Routing-slot construction

The implementation uses the following deterministic construction.

1. At each k-d tree recursion, station indices are sorted by the active coordinate. The lower half is processed before the upper half. The active coordinate alternates between latitude and longitude.
2. Indices inside each child are sorted in ascending input-index order before the next recursion. Leaves are returned in recursive lower-child-first order.
3. Each leaf preserves ascending input indices. If a leaf has fewer positions than the base spatial patch size, context indices are appended. Fixed-station experiments use affinity estimated from training observations. Spatial holdout uses haversine distance only.
4. Padded leaves are concatenated. Consecutive groups of leaves are reshaped into effective partitions according to the merge factor. Their concatenated positions become routing slots.
5. The recovery map selects original station positions, restores the source node order, and discards outputs at appended context positions.

Reindexing the same stations can change the within-leaf slot order and therefore the cross-partition routes. A permutation-invariant slot rule is outside the present study.

A.3 Spatial blocks and decoder

The implementation permits multi-head attention, while the reported experiments use one attention head. Query, key, and value projections include bias. Attention dropout and projection dropout are 0.1. Layer normalization has no affine parameters. The residual feed-forward block uses GELU, expansion ratio one, and dropout 0.1. Each layer computes the two spatial paths in parallel. The decoder reshapes the two temporal-patch representations and applies a shared one-by-one convolution to generate all forecast lead times.

For the global configuration, the tree produces 2,048 base leaves. A merge factor F gives $P = 2048/F$ effective partitions. The default $F = 32$ gives 64 partitions. Merge factors 8, 16, 32, and 64 are evaluated with the remaining backbone fixed.

B Experimental Details

B.1 Datasets and tasks

Table 7: Datasets and forecasting tasks.

Dataset	Spatial support and period	Nodes	Input	Output	Resolution
Global Wind	Physical stations worldwide, 2019 to 2020	3,850	48	24	Hourly
Global Temperature	Physical stations worldwide, 2019 to 2020	3,850	48	24	Hourly
MeteoNet NW	Physical rain gauges, 2016 to 2018	231	48	24	Hourly
MeteoNet SE	Physical rain gauges, 2016 to 2018	451	48	24	Hourly
Rain-UK	Regular one-degree precipitation mesh, 1979 to 2023	fixed mesh	30	30	Daily
Rain-USA	Regular one-degree precipitation mesh, 1979 to 2023	fixed mesh	30	30	Daily
Rain-CN	Regular one-degree precipitation mesh, 1979 to 2023	1,800	30	30	Daily
Rain-AUS	Regular one-degree precipitation mesh, 1979 to 2023	fixed mesh	30	30	Daily

The global station arrays contain 17,544 hourly steps from 2019-01-01 00:00 to 2020-12-31 23:00. Training uses 12,281 steps through 2020-05-26 16:00. Validation uses 1,754 steps through 2020-08-07 19:00. Testing uses the final 3,509 steps. Stored wind speed and temperature are converted back to physical units before reporting.

MeteoNet uses complete calendar years. Training, validation, and testing use 2016, 2017, and 2018. Six-minute precipitation reports are aggregated to an hourly value only when all ten reports are present. Stations require at least 80 percent completeness in the training year. Overall missingness is 4.27 percent for NW and 3.51 percent for SE. The missing rate rises from 1.79 percent to 6.45 percent between the NW training and test years, and from 1.60 percent to 5.11 percent for SE (Météo-France, 2020).

The gridded precipitation record contains 16,436 daily steps from 1979-01-01 to 2023-12-31. Training uses 11,505 days through 2010-07-02, validation uses 1,644 days through 2014-12-31, and testing uses 3,287 days from 2015 to 2023. These nodes are regular mesh cells and are not interpreted as physical rain gauges.

B.2 Evaluation protocols

Fixed-station forecasting uses recurring station identities and chronological time splits. Random-station holdout excludes 20 percent of stations from parameter optimization. Contiguous-region holdout projects coordinates onto a seed-dependent two-dimensional axis, uses one tail for spatial validation, uses the opposite 20 percent tail for testing, and uses the remaining stations for training. Held-out histories and labels are absent during optimization. Their recent histories are supplied at inference. Context completion for every spatial holdout uses geographic distance only.

The global record covers only two years, so its temporal robustness study uses rolling chronological blocks. The daily precipitation datasets support year-wise evaluation. For each test year from 2015 to 2023, the model is trained on data before the validation year, validated on the preceding year, and tested on the full target year. No input or target window crosses a calendar-year boundary.

B.3 Optimization, metrics, and uncertainty

The model uses Huber loss, AdamW, learning-rate scheduling, and gradient clipping at five. Global tasks use initial learning rate 0.002 and weight decay 0.001. Precipitation tasks use initial learning rate 0.002 and weight decay 0.0001. Training lasts at most 20 epochs.

For the global tasks, validation searches spatial-layer counts in $\{3, 4, 5, 6, 8\}$, temporal patch lengths in $\{6, 12, 15, 24, 30\}$, and base spatial patch sizes in $\{2, 4, 8\}$. The selected values are five layers, 24 temporal steps, and two stations. For the daily precipitation tasks, the searches use $\{3, 4, 5, 6\}$ layers, temporal patch lengths $\{5, 10, 15, 30\}$, and spatial patch sizes $\{2, 4, 8\}$. The selected values are five, 15, and two.

Continuous forecasts are evaluated with mean absolute error, mean squared error, and root mean squared error. Precipitation event evaluation uses probability of detection, false-alarm ratio, critical success index, equitable threat score, Heidke skill score, and Brier score. Thresholds include 0.1, 1, 5, and 10 millimetres when appropriate.

Controlled studies use seeds 1 through 5. Confidence intervals are

$$\bar{m} \pm t_{4,0.975} \frac{s_m}{\sqrt{5}}. \tag{15}$$

Paired seed analyses compare matched runs. Conditional temporal intervals use 5,000 bootstrap repetitions and resample complete non-overlapping blocks of 168 forecast origins within each seed. The final incomplete block is excluded. Seed labels remain fixed, and resampled seed means are averaged before percentile limits are formed. Controlled runs use PyTorch 2.11.0, CUDA 12.8, and one RTX 5080. A separate engineering profile uses one RTX 4090.

B.4 Baselines and data provenance

The comparison includes persistence, seasonal lag, station climatology, ARIMA, Holt-Winters, gradient-boosted trees, Prophet, GFS, ERA5, DeepAR, N-BEATS, Pyraformer, Autoformer, In-

former, CATS, TimeMixer, StemGNN, Corrformer, MRIformer, DLinear, PatchTST, and STELLA using its released implementation. Learning-based comparisons use matched input and output windows and the same temporal splits. The STELLA-equivalent implementation was checked against released code at commit `cdb64e2d4b61`. Both implementations have 9,976 parameters, and the maximum absolute forward difference after parameter mapping is 4.768×10^{-7} (**?**).

The global station arrays are attributed to the Integrated Surface Database (Smith et al., 2011). A metadata audit matches 3,688 of 3,850 coordinate pairs to the official station history after rounding to 0.001 degrees. It also matches elevation within one metre for 3,661 stations. MeteoNet SE is obtained from the official Météo-France distribution. MeteoNet NW is obtained from the public MeteoNet distribution. Acquisition routes, processing summaries, checksums, split indices, and run manifests are recorded in the experiment archive.

C Complete Fixed-Station Results

C.1 Controlled five-seed metrics

Table 8: Controlled fixed-station forecasting on one RTX 5080. Entries are five-seed means with 95 percent t intervals. Delta MAE is comparison minus WeatherScaleNet.

Dataset	Model	MAE	RMSE	MSE	Delta MAE and seed CI	Block 95 percent CI	Holm p_t
Global Wind	WeatherScaleNet	1.2066 ± 0.0023	1.8030 ± 0.0038	3.2507 ± 0.0138	0	0	not applicable
Global Wind	STELLA	1.3174 ± 0.0002	1.9660 ± 0.0015	3.8653 ± 0.0058	0.1108 [0.1085, 0.1131]	[0.1069, 0.1149]	< 0.001
Global Wind	TimeMixer	1.3271 ± 0.0041	1.9619 ± 0.0035	3.8489 ± 0.0137	0.1206 [0.1151, 0.1260]	[0.1159, 0.1232]	< 0.001
Global Temperature	WeatherScaleNet	1.5944 ± 0.0066	2.4080 ± 0.0102	5.7984 ± 0.0491	0	0	not applicable
Global Temperature	STELLA	1.9310 ± 0.0018	2.8682 ± 0.0017	8.2268 ± 0.0095	0.3366 [0.3306, 0.3427]	[0.3245, 0.3496]	< 0.001
Global Temperature	TimeMixer	1.9005 ± 0.0021	2.8172 ± 0.0031	7.9367 ± 0.0175	0.3061 [0.2992, 0.3129]	[0.2949, 0.3201]	< 0.001

C.2 Broader benchmark comparison

Table 4 compares the model families evaluated in the single-run global benchmark. GFS and ERA5 are external references rather than WeatherScaleNet inputs. Reanalysis can incorporate observations unavailable to an operational forecaster. The single-run results and controlled five-seed results are kept separate.

WeatherScaleNet requires more computation than gradient-boosted trees, DLinear, and STELLA. The accuracy and resource comparisons address different model-selection criteria.

C.3 Lead time, terrain, and regional error

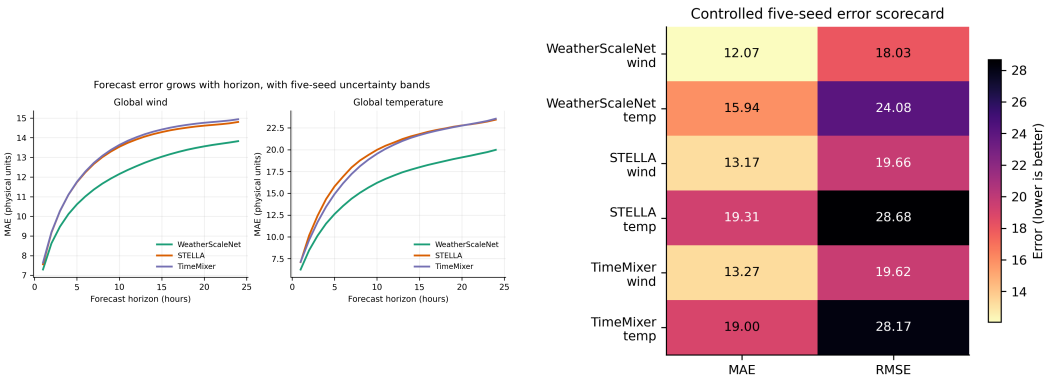

Figure 2: Controlled fixed-station results. The left panel reports lead-wise mean absolute error with five-seed intervals. The right panel summarizes aggregate error metrics.

The terrain analysis divides the 3,850 wind stations into plains below 200 metres, hills from 200 to 500 metres, and mountains above 500 metres. The largest displayed gain occurs in hilly and

mountainous regions. At the 12-hour lead for mountain stations, the absolute error reduction relative to Corrformer is 15.6 percent.

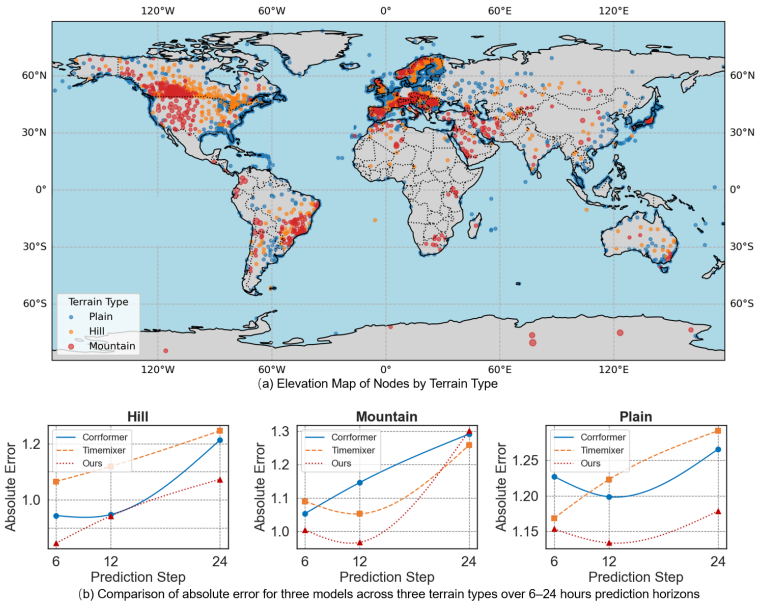

Figure 3: Station terrain classes and lead-wise mean absolute error for the global wind task.

The regional diagnostic covers the eastern Mediterranean, Balkans, Anatolia, and the Black Sea to Caucasus sector. The displayed lead times are 6, 12, and 24 hours. WeatherScaleNet shows fewer concentrated high-error regions than TimeMixer and Corrformer in the displayed period.

D Held-Out Station and Rain-Gauge Diagnostics

D.1 Representative Spatial Splits

D.2 Physical Rain Gauges and Native Missingness

Table 9: Hourly precipitation forecasting on physical MeteoNet gauges. Mean absolute error is the five-seed mean with a 95 percent t interval. Critical success index is computed at hourly thresholds of 0.1 and 1.0 millimetres.

Region	Evaluation	Representation	MAE	MSE	CSI at 0.1	CSI at 1.0
NW	All gauges	Pretrained coordinates	0.1195 ± 0.0034	0.2666	0.1556	0.0138
NW	Random holdout	Pretrained coordinates	0.1174 ± 0.0036	0.2598	0.1615	0.0146
NW	Block holdout	Pretrained coordinates	0.1168 ± 0.0068	0.2427	0.1484	0.0087
SE	All gauges	Pretrained coordinates	0.1624 ± 0.0073	1.8191	0.1958	0.0458
SE	Random holdout	Pretrained coordinates	0.1639 ± 0.0068	0.6526	0.1923	0.0473
SE	Block holdout	Pretrained coordinates	0.1685 ± 0.0192	2.4320	0.1823	0.0339

Average error remains low under both spatial holdouts, but event scores decrease as the precipitation threshold increases. Critical success index is approximately zero at 5 millimetres per hour under the standard Huber objective. Predicted variance is below observed variance, which indicates smoothing of heavy rainfall. These results establish forecasting feasibility for location-known gauges with recent observations. They do not establish strong heavy-rain detection.

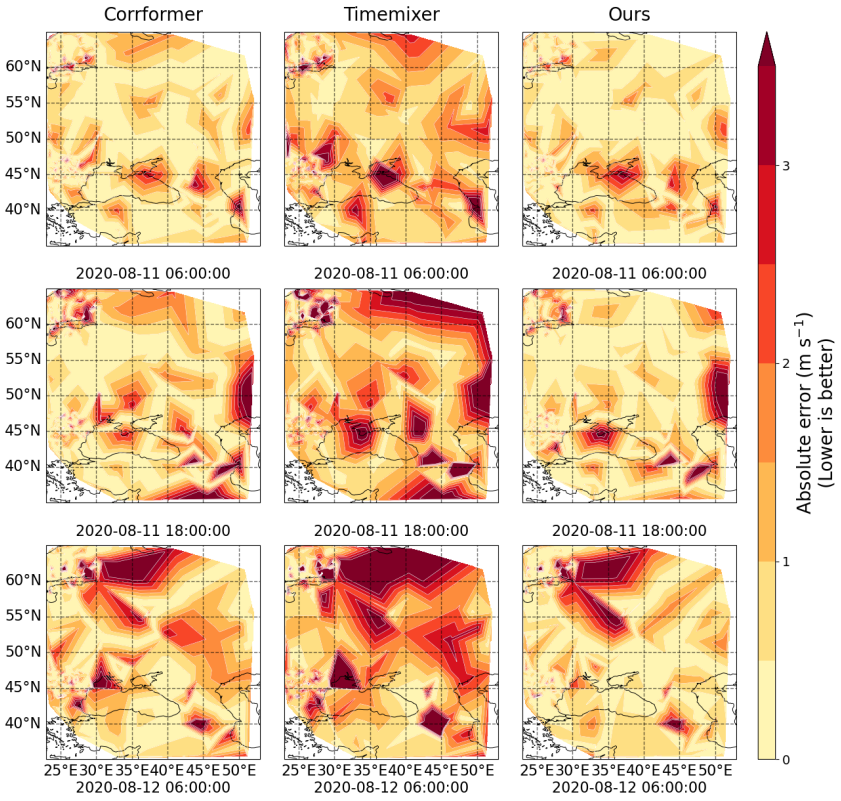

Figure 4: Regional wind forecast error at 6, 12, and 24-hour leads.

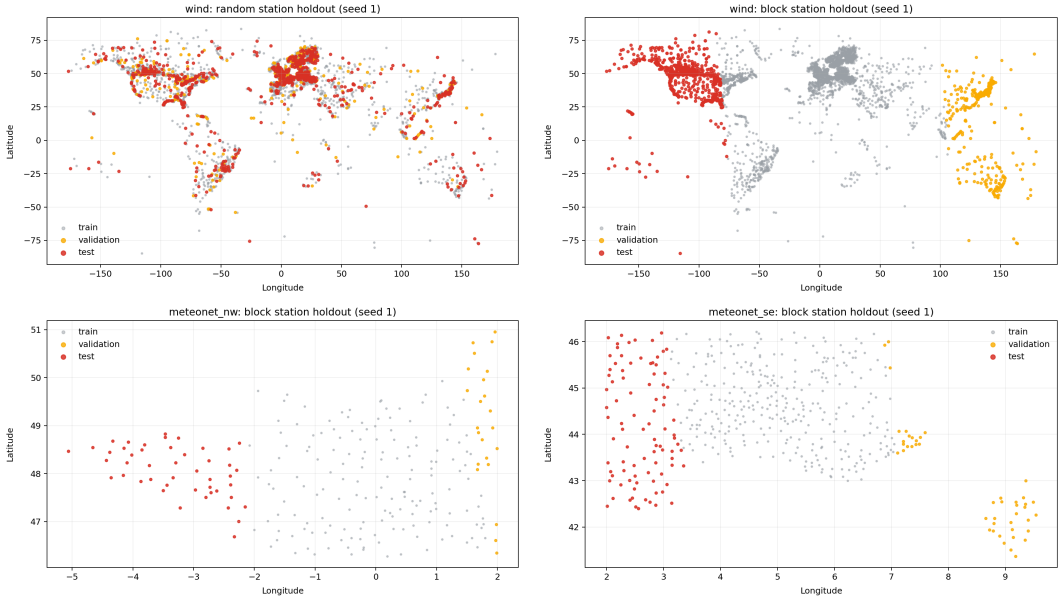

Figure 5: Representative spatial holdouts. The panels show global random holdout, global contiguous-region holdout, MeteoNet NW contiguous-region holdout, and MeteoNet SE contiguous-region holdout.

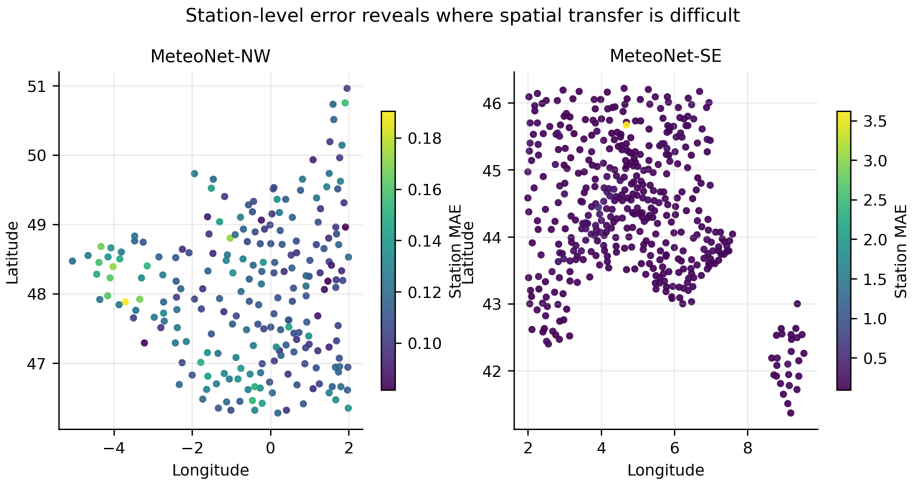

Figure 6: Station-level test mean absolute error on the MeteoNet physical rain-gauge networks. Each point represents one gauge.

E Communication and Factorial Controls

E.1 Communication-axis experiment

The fixed-station branch experiment compares a within-partition-only model, a cross-partition-only model, and a model retaining both paths. Both single-path models have higher five-seed mean absolute error on Global Wind and Global Temperature. Paired t analyses support this direction. Exact Wilcoxon tests do not remain significant after Holm correction with five seed pairs. The branch summary records the ordering and paired-test direction but does not provide run-level means or intervals, so no numerical effect size is reported.

E.2 Grouped and factorial ablations

The grouped ablation removes the history and geospatial station context, the normalized learned aggregation, or both grouped choices. It records the behavior of the configured system but does not identify which element inside a group is responsible for the change.

Table 10: Grouped ablation across Global Wind, Global Temperature, and Rain-CN.

Configuration	Observed result
Full configuration	Reference for the grouped comparison
Without coordinate and elevation station context	Error increases on all three reported targets
Without normalized learned aggregation	Error increases on all three reported targets
Without both grouped changes	Largest degradation in the grouped comparison

The factorial study separates history, coordinate representation, elevation, the two spatial paths, and the aggregation rule. Position information helps in fixed-station global controls, but the winning coordinate representation changes across held-out settings. Elevation helps both global targets, with quantile bins slightly stronger for wind and a continuous elevation network slightly stronger for temperature. No measured scalar aggregation layer collapses completely to one path. Feature concatenation gives the lowest global error among the tested aggregation rules, while normalized scalar aggregation has fewer parameters and preserves the representation dimension.

Table 11: Evidence and interpretation for the principal controls.

Tested element	Empirical finding	Interpretation
Shared representation without an untrained identity lookup	Shared weights generate the held-out token; no identity-indexed trainable row left untrained by the split is required.	Learned-computation boundary
Protocol-respecting route construction	WeatherScaleNet uses static coordinates and geographic distance. No route statistic is estimated from held-out training observations.	Information boundary
Within-partition and cross-partition paths	Either single path is worse than both paths on the two fixed-station global controls.	Communication result in the studied family
Coordinate representation	Pretrained coordinates, raw coordinates, Fourier features, and no position have different winners.	Replaceable implementation
Elevation representation	Quantile bins and a continuous network exchange rank across targets.	Replaceable implementation
Aggregation rule	Concatenation improves accuracy, while normalized scalars reduce parameters and expose path weights.	Replaceable implementation
Partition count	Error changes little across the tested merge factors.	Replaceable in the tested range

F Partition, Missingness, and Temporal Robustness

F.1 Partition scale and boundary diagnostics

The default global layout starts from 2,048 base leaves and joins consecutive leaves with merge factor 32, giving 64 effective partitions. Merge factors 8, 16, 32, and 64 are evaluated with the remaining backbone fixed. Global mean absolute error changes by less than 0.01 for wind and by about 0.011 for temperature. The absolute correlation between station error and distance to the nearest partition boundary is at most 0.091.

Table 12: Partition-scale and boundary diagnostics.

Diagnostic	Tested range	Result
Merge factor	8, 16, 32, 64	Effective partition counts 256, 128, 64, 32
Global Wind MAE span	Four layouts	Less than 0.01
Global Temperature MAE span	Four layouts	About 0.011
Absolute boundary distance-error correlation	All tested layouts and targets	At most 0.091

The results do not identify one privileged partition count and do not show a systematic aggregate boundary penalty. They do not establish station-level invariance to a boundary.

F.2 Missing-input shift

Matched missingness during training and testing degrades performance gradually. Missingness introduced only at test time has a larger effect, especially at 30 to 50 percent corruption for temperature. The observation mask prevents missing targets from contributing to the loss and metric denominator. It does not guarantee robustness to a missingness regime absent from training.

F.3 Rolling temporal evaluation

The global record uses rolling chronological blocks rather than nine held-out years. The daily precipitation datasets use literal year-wise evaluation from 2015 to 2023. This experiment tests temporal shift separately from station-identity holdout.

Table 13: Missing-input experiments.

Protocol	Result
Matched corruption during training and testing	Error increases gradually as the missing fraction rises
Test-only corruption	Larger degradation than matched corruption
Test-only corruption from 30 to 50 percent	Strongest degradation on Global Temperature
Native MeteoNet shift	Test missingness exceeds training missingness in both regions

G Regional Gridded Precipitation and Event Skill

G.1 Daily precipitation benchmarks

The four regional datasets test whether the coordinate-indexed implementation applies across regions and mesh geometries. They are regular-mesh fields rather than physical rain-gauge networks. Table 14 reports the complete benchmark. WeatherScaleNet has the lowest mean absolute error on all four datasets. It also has the lowest mean squared error and root mean squared error on Rain-AUS, Rain-CN, and Rain-USA. DLinear has the lowest squared-error metrics on Rain-UK, while WeatherScaleNet retains the lowest mean absolute error.

Table 14: Regional daily precipitation benchmark under the established single-run protocol. The input and forecast windows are both 30 days. Lower is better.

Method	Rain-AUS			Rain-CN			Rain-UK			Rain-USA		
	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Graph-based models												
Graph WaveNet	2.098	10.246	3.201	4.544	61.351	7.835	3.229	21.427	4.629	3.184	50.084	7.077
STGCN	2.107	10.163	3.188	4.555	61.231	7.825	3.245	21.409	4.627	3.197	50.240	7.088
StemGNN	2.488	13.250	3.640	5.126	51.094	7.148	3.775	26.853	5.182	2.118	23.445	4.842
Temporal forecasting models												
FEDformer	2.171	10.183	3.191	4.723	62.252	7.890	3.376	21.707	4.659	3.348	51.998	7.211
Historical inertia	2.768	15.023	3.876	5.888	78.996	8.888	4.026	30.019	5.479	4.102	66.830	8.175
PatchTST	2.160	10.336	3.215	4.655	64.272	8.017	3.316	21.772	4.666	3.321	53.685	7.327
DLinear	2.135	10.310	3.211	4.672	63.696	7.981	3.320	**21.077**	**4.591**	3.282	54.228	7.364
ConvLSTM	2.110	10.323	3.213	4.566	61.717	7.856	3.266	21.576	4.645	3.225	53.509	7.315
DeepAR	2.450	12.800	3.578	5.210	68.450	8.273	3.650	25.400	5.040	3.750	58.200	7.629
N-BEATS	1.910	10.150	3.186	3.320	46.850	6.845	2.940	22.450	4.738	1.480	22.100	4.701
TimesNet	1.835	9.998	3.162	3.152	44.582	6.677	2.816	21.996	4.690	1.371	21.279	4.613
Autoformer	2.631	15.960	3.995	5.407	75.690	8.700	4.378	39.362	6.274	2.420	26.255	5.124
CATS	1.858	10.131	3.183	3.051	44.783	6.692	2.891	22.524	4.746	1.354	22.297	4.722
TimeMixer	1.835	10.093	3.177	3.064	44.435	6.666	2.858	24.572	4.957	1.367	22.543	4.748
Informer	1.825	9.935	3.152	3.025	45.279	6.729	2.843	23.804	4.879	1.369	21.893	4.679
Pyraformer	2.240	10.550	3.248	4.850	63.800	7.987	3.420	22.150	4.706	2.950	51.200	7.155
Meteorological forecasting models												
Corrformer	1.843	10.446	3.232	3.113	46.077	6.788	2.881	22.552	4.749	1.320	21.344	4.620
MRIformer	1.950	10.650	3.263	3.450	48.200	6.943	2.990	23.100	4.806	1.450	21.800	4.669
STELLA	2.045	12.159	3.487	3.081	49.952	7.068	2.885	25.294	5.029	1.351	23.590	4.857
WeatherScaleNet	**1.821**	**9.820**	**3.134**	**3.017**	**44.409**	**6.664**	**2.773**	22.315	4.724	**1.306**	**21.252**	**4.610**

General-purpose sequence models can match or exceed Corrformer in several rain settings. This result indicates that target statistics affect the value of long-range station coupling. The table is used for fixed-node benchmark coverage and does not serve as evidence about physical rain-gauge transfer.

G.2 Precipitation event trade-off

Low average precipitation error can coexist with weak detection of rare events. The Rain-UK loss control quantifies the trade-off between average error and event recall. Increasing the event weight improves critical success index at 5 millimetres, but mean absolute error and false alarms also increase.

Table 15: Rain-UK event-weighted loss control at the 5 millimetre threshold.

Training objective	CSI at 5 mm	MAE
Standard Huber	approximately 0.001	2.77
Event weight 4	0.143	3.76
Event weight 8	0.220	4.86

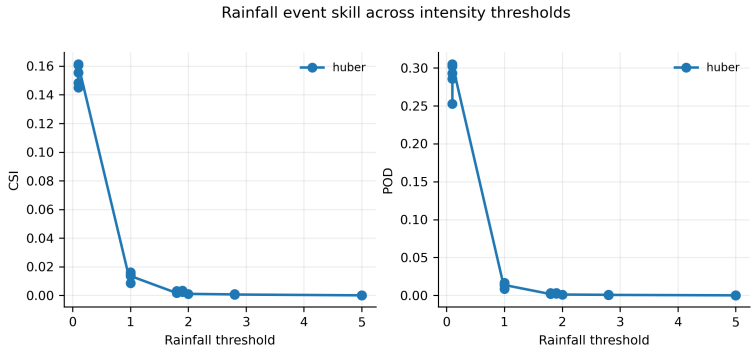

Figure 7: Precipitation event skill across intensity thresholds and loss settings.

The loss study does not identify one uniformly better objective. Event detection, false alarms, and average error must be selected according to the application.

H Computational Efficiency

Table 16: Controlled end-to-end efficiency on one RTX 5080 with batch size eight. Values are five-seed means. Throughput is complete forecast origins per second.

Dataset	Model	Parameters	Peak memory, MB	Train epoch, s	Test origins per s
Global Wind	WeatherScaleNet	1,016,514	2675.5	49.85	679.26
Global Wind	STELLA	9,976	102.6	4.97	6400.52
Global Wind	TimeMixer	57,665	6694.1	237.75	217.17
Global Temperature	WeatherScaleNet	818,368	2174.3	40.56	820.71
Global Temperature	STELLA	9,976	102.5	4.98	6569.12
Global Temperature	TimeMixer	57,665	6694.1	230.24	221.08

WeatherScaleNet is slower and larger than STELLA. It uses less peak memory and has higher test throughput than TimeMixer in the matched comparison. A separate RTX 4090 engineering profile records about 4.4 GB peak training memory for WeatherScaleNet, 6.7 GB for TimeMixer, and 8.8 GB for PatchTST on Global Wind. It records 21.1 milliseconds per inference iteration for WeatherScaleNet and 34.1 milliseconds for TimeMixer. The two hardware settings are not combined in one quantitative comparison.

The measurements do not establish deployment suitability. Operational use also depends on preprocessing, data transfer, reliability requirements, and target hardware.

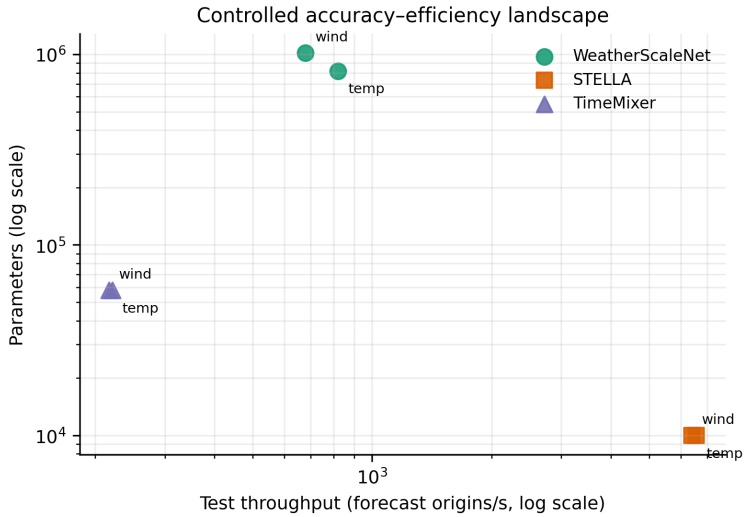

Figure 8: Accuracy and efficiency under the matched hardware comparison.