CyPhy-STN: Cycle-aware Physics-inspired Message Passing for Multivariate Spatio-Temporal Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Can a single physics-inspired architecture perform consistently across diverse time-series prediction tasks governed by different dynamics? We introduce the Cycle-aware Physics-inspired Spatio-Temporal Network (CyPhy-STN), a unified architecture for multivariate spatio-temporal forecasting. It combines cyclic temporal attention with parallel, mechanism-specific graph propagation inspired by continuity, viscosity, relaxation, advection, and reaction. Learnt channel- and regime-aware routing determines which physical variables enter each mechanism and how messages are propagated, allowing the same architecture to be applied across different domains without the need of a fixed governing equation or an architectural redesign. Across four traffic and atmospheric forecasting benchmarks, CyPhy-STN achieves lower cross-benchmark regret and average rank than state-of-the-art spatio-temporal models with a low parameter count.

1 Introduction

Many real-world physical systems unfold simultaneously across three dimensions: time, space, and different features as characteristics. Traffic networks, for example, consist of measurements such as flow, occupancy and speed at multiple locations over time, while atmospheric systems combine pollution concentrations and meteorological variables across spatially distributed stations. Multivariate spatio-temporal forecasting tries to model dependencies along all three dimensions.

This raises as a central question: How should dependencies be modelled when observations vary simultaneously across space, time, and—possibly coupled—physical quantities? Existing approaches typically focus their inductive biases on particular parts of this dependency structure. Current state-of-the-art (SotA) forecasting approaches are predominantly data-driven, such as spatio-temporal graph neural networks (ST-GNNs) that combine temporal modelling with spatial message passing or attention-based approaches that learn temporal or cross-variable dependencies. In addition to purely data-driven methods, physics-informed methods provide an additional source of inductive bias by incorporating known physical structure and dynamics. However, it is difficult to decide which interactions across space, time, and feature dimensions should be learnt freely and which should be separated or constrained by physical considerations. This often results in specialised models that require substantial expert knowledge and costly adaptation to transfer across tasks.

We propose the **Cycle-aware Physics-inspired Spatio-Temporal Network (CyPhy-STN)** as a unified architecture that is based on a structured decomposition of these three dimensions. In place of treating all three symmetrically, CyPhy-STN uses different inductive biases for each type of dependency. First, within each physical channel, Temporal Cyclic Multi-Head Attention (TC-MHA) learns relationships between the temporal trajectories of nodes in a data-driven fashion which extends the cyclic querying mechanism of TQNet (Lin et al., 2025) to multi-variate graph time series. Secondly, the resulting node- and channel-specific representations are then propagated along the physical graph, where spatial and cross-variable interactions are structured and enriched by physics-inspired message passing. Rather than learning a single black-box message function or constraining learning to a fixed governing PDE, CyPhy-STN decomposes physical interactions into a suite of interpretable, physical transport mechanisms, namely, continuity, viscosity, relaxation, advection, and reaction. Analytical physical formulations are retained analytically where possible, while the unknown or system-specific dynamics are learnt using lightweight Multi-Layer Perceptrons (MLPs).

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.

Unlike operator-splitting approaches that apply physical mechanisms sequentially (Eliasof et al., 2024), CyPhy-STN evaluates all mechanisms in parallel on the same graph state and the contributions are adaptively combined through learnt edge-wise gates.
A channel- and regime-aware routing scheme determines which physical quantities enter each mechanism and to which channels its messages contribute. The mechanisms stay domain-independent, this means advection is the same operator for traffic flow as for PM2.5 in atmospheric datasets. The routing specifies which measured channels play which physical role in a given system. This allows the same architecture to be instantiated across structurally different systems without prescribing a single governing equation.
We evaluate CyPhy-STN on four benchmarks, that cover highway traffic and atmospheric forecasting. In contrast to common single-target settings, the model jointly forecasts all observed physical variables at all nodes. CyPhy-STN achieves consistently competitive performance even compared to specialised models and shows low cross-benchmark regret, demonstrating strong adaptability across diverse forecasting tasks. Ablations demonstrate complementary contributions of its temporal and physics-based pathways. Our main contributions are:

- We introduce a **structured decomposition of multi-variate spatio-temporal dependencies**, using different inductive biases for temporal, spatial, and cross-variable interactions rather than trying to learn unrestricted interactions across all three dimensions.
- We propose **parallel, routed physics message passing**: analytically specified and learnt mechanisms are adaptively combined rather than a fixed governing PDE, with each mechanism routed between designated source and target channels, enabling one architecture to span structurally different physical domains.
- Across traffic and atmospheric forecasting, CyPhy-STN shows **consistent cross-domain performance** with a comparatively low parameter count. Ablations demonstrate complementary, domain-dependent contributions of its temporal and physics-based pathways.

2 Related Work

Purely Data-Driven Spatio-Temporal Models Purely data-driven spatio-temporal models learn spatial and temporal dependencies directly from observations, using architectural inductive biases such as graph structure, attention, or representation learning rather than explicit governing physical principles. Hosseini et al. (2026) proposed HG-GFNO for traffic forecasting, combining static and adaptive graph convolutions with a Graph Fourier Neural Operator. Its frequency-domain HGFO block captures global node dependencies without the $\mathcal{O}(L^2)$ complexity of full attention. Spatio-Temporal Adaptive Embedding Transformer (STAEformer) by Liu et al. (2023) is a spatio-temporal traffic forecasting model which combines static spatial and temporal embedding with adaptive embeddings to learn hidden spatio-temporal structure. Spatial-Temporal-Decoupled Masked Pre-training for Spatiotemporal Forecasting (STD-MAE) by Gao et al. (2024) employs separate spatial and temporal masked autoencoders to learn long-range, context-rich representations, which are subsequently integrated into arbitrary downstream spatio-temporal predictors. The Dynamic Trend Representation Transformer (DTRformer) by Chen et al. (2025) extends the STAEformer architecture by introducing explicit spatial–temporal interaction through cross-attention and incorporating predefined road-network information through dynamic–static graph fusion. MSTHH (Wang et al., 2026) formulates traffic forecasting as a multi-modal task, jointly encoding co-located flow, speed, and occupancy measurements as input features to exploit cross-modal dependencies.

Physics-Informed and Physics-Inspired GNNs The incorporation of physics into spatio-temporal models can be put to two distinct categories: physics as training objective and physics as propagation operator or architecture inductive biases. The first approach was employed by Shi et al. (2021) who combine a Physics-Uninformed Neural Network for state estimation with a Physics-Informed Neural Network that regularises the predictions using a second-order traffic-flow model. Similarly, Assiotis et al. (2025) propose a graph-convolutional gated recurrent unit with a PDE-based physics loss for air-quality forecasting. Representative works for the latter are GRAND by Chamberlain et al. (2021) which interprets graph representation learning as a continuous diffusion process and connects GNN propagation with the heat equation and the Advection-Diffusion-Reaction GNN (ADR-GNN) by Eliasof et al. (2024) which extends this perspective from diffusion to multiple transport mechanisms, implemented through operator-split graph updates with a mass-preserving advec-

Table 1: Physics-structured ST models, **C**ontinuity, **D**iffusion, **A**dvection, **R**eaction, rela**X**ation.

Model	Mechanisms	Composition	Routing
GRAND Chamberlain et al. (2021)	D	single operator	—
SSTODE Jiang et al. (2026)	DA	fixed operator set	—
AirDDE Wu et al. (2026)	DAR	fixed set + delays	—
Air-DualODE Tian et al. (2025)	DAR	dual branch	—
ADR-GNN Eliasof et al. (2024)	DAR	operator-split	—
CyPhy-STN (ours)	CDARX	parallel, gated	channel-wise

tion operator. Air-DualODE (Tian et al., 2025) combines a physics-guided open-system ODE with a data-driven Neural ODE for air-quality prediction. Its physical branch models advection, diffusion, dissipation, and generation, while an additional attention-based latent dynamics model captures residual spatio-temporal dependencies. SSTODE by Jiang et al. (2026) models sea-surface temperature dynamics with a continuous-time Neural ODE parameterizing advection–diffusion transport, while reconstructing latent velocity fields from sparse observations and accounting for external heat fluxes. With AirDDE, Wu et al. (2026) propose an air-quality forecasting model based on Neural Delay Differential Equations (NDDEs) using diffusion, delayed advection and a source/sink term. These approaches encode domain-specific physical dynamics through continuous-time differential equations or predefined combinations of physical processes. As a result, they are often specialised to a particular domain and require expert-driven adaptation when applied to different tasks. An overview is given in Table 1. CyPhy-STN, however, represents physical mechanisms as parallel, mechanism-specific graph updates with adaptive contributions, enabling different physical processes and variable roles to be composed without assuming a single fixed governing equation.

3 METHOD

3.1 PROBLEM FORMULATION

We consider multivariate spatio-temporal observations along three dimensions: space, physical variables, and time. Let $\mathbf{X} \in \mathbb{R}^{N \times C \times T_{\text{in}}}$ denote the historical spatio-temporal tensor, where N is the number of spatial nodes, C the number of physical variable channels, and T_{in} the temporal look-back window. Our objective is to learn a mapping f that jointly forecasts the future evolution of all C channels across all N nodes over a prediction horizon T_{out}: $f : \mathbb{R}^{N \times C \times T_{\text{in}}} \to \mathbb{R}^{N \times C \times T_{\text{out}}}$. This extends the common single-variable forecasting formulation, in which the target is restricted to $\mathbb{R}^{N \times T_{\text{out}}}$, to joint prediction across the variable dimension.

3.2 MODEL OVERVIEW

An overview of the proposed **CyPhy-STN** architecture is shown in Figure 1. CyPhy-STN explicitly preserves the three dimensions of the prediction problem—time, space, and physical variables—while applying different inductive biases to their interactions. The input $\mathbf{X}_{\text{in}}$ is first normalised via Reversible Instance normalisation (RevIN). Temporal Cyclic Multi-Head Attention (TC-MHA) then processes each physical channel independently: nodes form the attention tokens and their temporal trajectories the corresponding representations, allowing temporal patterns and cyclic dependencies to inform node-to-node interactions without mixing heterogeneous physical variables. The resulting representations are projected to the model dimension d and augmented with a learnt node embedding $\mathbf{E}_{\text{node}}$ and a input context embedding $\mathbf{E}_{\text{context}}$. A linear integration layer maps these components to the hidden state $\mathbf{H}_{\text{proj}} \in \mathbb{R}^{N \times C \times d}$ which retains separate representations for every node and physical channel. A stack of Physics Message Passing layers subsequently models spatial and cross-variable interactions. Using the physical graph topology, edge attributes, cyclic embeddings $c_t = \mathbf{E}p\big[, \text{idx}_{\text{cyc}} \bmod L, \big]$, and statistics derived from the raw input $\mathbf{X}_{\text{in}}$, information is propagated through parallel, mechanism-specific messages inspired by continuity, viscosity, relaxation, advection, and reaction. Channel-aware routing determines which physical variables are involved in which mechanism and to which channels its messages contribute. A prediction head maps the resulting latent state to the forecasting horizon, followed by the inverse RevIN transformation.

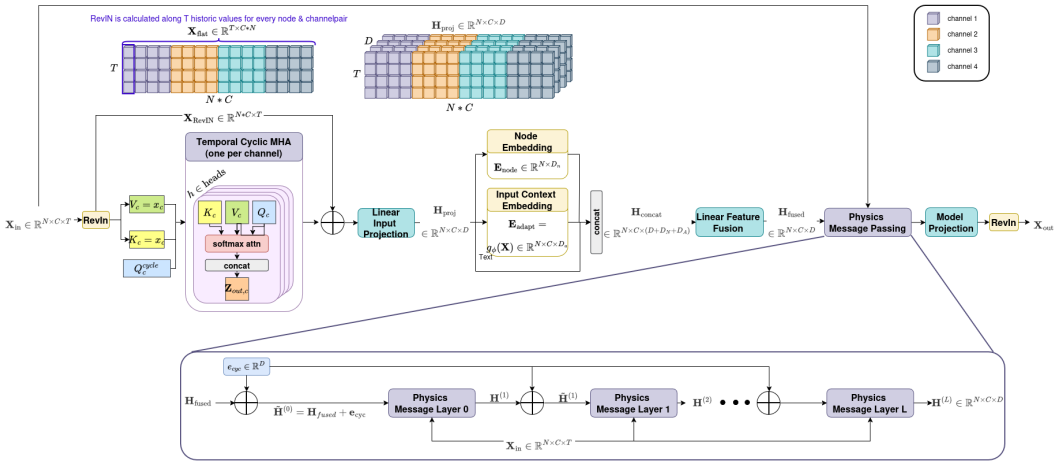


Figure 1: CyPhy-STN architecture with per-batch tensor shapes (T: time sequence, N: nodes, C: channels, d: hidden dimension).

3.2.1 Temporal Cyclic Multi-Head Attention (TC-MHA)

To capture macro-temporal dependencies while preserving channel independence, we adapt the cross-variable periodic querying of TQNet Lin et al. (2025) to graphs by treating spatial nodes as attention tokens. Each token summarises a node's T-step history, so attention relates node trajectories rather than individual time steps. The mechanism is applied separately to each channel $c \in \{1, \dots, C\}$: temporal histories determine inter-node relationships, while heterogeneous physical variables are never mixed. For each channel, a cyclic query is built from a learnable periodic embedding $\mathbf{E}_c \in \mathbb{R}^{P \times N}$, indexed by the phases of the input window:

$$\begin{aligned} \mathbf{Q}_c &= \left[\mathbf{E}_c[(t+\tau) \bmod P]\right]_{\tau=0}^{T-1\top} \in \mathbb{R}^{N \times T}, \\ \mathbf{Z}_c &= \mathrm{MHA}_c(\mathbf{Q}_c, \mathbf{X}_c, \mathbf{X}_c) \in \mathbb{R}^{N \times T}, \\ \mathbf{Z}_{\text{out}} &= \mathrm{Concat}\left(\{\mathbf{Z}_c\}_{c=1}^{C}\right) \in \mathbb{R}^{(N \cdot C) \times T_{\text{in}}}. \end{aligned} \tag{1}$$

With nodes as tokens and their histories as features, each of the H heads yields a channel-specific $N \times N$ affinity matrix that conditions node-to-node aggregation on cyclic position. Cross-channel interactions are deliberately deferred to the physics-inspired message-passing stage, where they are structured by physical roles rather than learnt as unrestricted attention.

3.2.2 Input Projection and Embeddings

Before physics-based message passing, the explicit temporal axis is collapsed into one latent state per node and channel. $\mathbf{Z}_{\text{out}}$ is projected to the model dimension d:

$$\mathbf{Z}_{\text{proj}} = \mathbf{Z}_{\text{out}} \mathbf{W}_{\text{proj}}^{\top} \in \mathbb{R}^{(N \cdot C) \times d}, \quad \mathbf{W}_{\text{proj}} \in \mathbb{R}^{d \times T_{\text{in}}}, \tag{2}$$

so temporal information is retained in the latent features while the tensor shape changes from $N \times C \times T_{\text{in}}$ to $N \times C \times d$. We enrich this representation with two auxiliary embeddings:

1. **Static Node Embeddings:** a learnable matrix $\mathbf{E}_{\text{node}} \in \mathbb{R}^{N \times d_n}$ (Xavier-uniform initialisation), shared across samples and broadcast over channels, capturing time-invariant node characteristics.
2. **Input Context Embeddings:** a two-layer ReLU MLP $g_\phi : \mathbb{R}^{T_{\text{in}}} \to \mathbb{R}^{d_{\text{context}}} \to \mathbb{R}^{d_{\text{context}}}$ applied to each node–channel series yields $\mathbf{E}_{\text{context}} = g_\phi(\mathbf{X}_{\text{in}}) \in \mathbb{R}^{N \times C \times d_{\text{context}}}$. It captures sample-specific local context and, unlike $\mathbf{Z}_{\text{proj}}$, bypasses temporal attention.

After reshaping $\mathbf{Z}_{\text{proj}}$ to $N \times C \times d$, a linear projection gives the initial latent physics layer state:

$$\mathbf{H}_{\text{proj}} = \mathrm{Linear}\left(\left[\mathbf{Z}_{\text{proj}} \,\|\, \mathbf{E}_{\text{node}} \,\|\, \mathbf{E}_{\text{context}}\right]\right) \in \mathbb{R}^{N \times C \times d}, \tag{3}$$

where $\|$ denotes concatenation along the feature dimension. This preserves a separate state for each node and physical variable, on which the subsequent graph-based physical interactions operate.

3.2.3 PHYSICS-INSPIRED MESSAGE PASSING

The physics message-passing layers structure interactions across nodes, following the graph topology, and across variables, following the physical roles a routing map assigns within channel groups. For efficiency, $\mathbf{X}_{\text{phys}}$ is summarised by its temporal mean $\bar{\mathbf{X}}_{\text{phys}}$ and trend $\Delta\mathbf{X}_{\text{phys}} = \mathbf{X}_{\text{phys}}(T) - \mathbf{X}_{\text{phys}}(1)$, from which the routing map selects $(\bar{\rho}, \bar{v}, \bar{q}, \bar{\rho}^{\text{trend}})$ for each message function. We set $\mathbf{H}^{(0)} = \mathbf{H}_{\text{proj}}$ and add a learnt time-of-cycle embedding $\mathbf{p}_t$ at each layer. Rather than assuming a fixed PDE, we motivate the messages from generic transport dynamics for a density-like field ρ and a velocity-like field v:

$$\underbrace{\partial_t \rho + \partial_x(\rho v)}_{\text{continuity}} = \underbrace{S(\rho)}_{\text{reaction}}, \tag{4}$$

$$\partial_t v + \underbrace{\partial_x(uv)}_{\text{advection}} = \underbrace{\mu\, \partial_x^2 v}_{\text{diffusion}} + \underbrace{\frac{v_{\text{eq}}(\rho) - v}{\tau(\rho)}}_{\text{relaxation}}, \tag{5}$$

with transport velocity u learnt independently of v. Each of the five terms becomes a separate message function, specified analytically where the physics is known and learnt where it is not: continuity and diffusion use fixed discretisations, relaxation uses global learnt parameters, advection is learnt edge-wise, and reaction, which in environmental systems stems from unresolved chemical, micrometeorological, and emission processes, is learnt node-wise.

Channel Grouping Physical channels are organised into semantically coherent groups $g \in \mathcal{G}$, each corresponding to a distinct physical regime such as near-surface or lower-atmosphere dynamics. Within each group, if there are two velocity component channels, they are assembled into a nodal vector field $\bar{\mathbf{v}}_i^{(g)} \in \mathbb{R}^2$, which drives the transport of the group's scalar channels. Since the edge-level transport messages require a scalar velocity along each edge, we project the nodal vector fields onto the unit edge direction $\mathbf{e}_{ij}$ (positive values indicating flow from i toward j):

$$\bar{v}_i^{(g,ij)} = \bar{\mathbf{v}}_i^{(g)} \cdot \mathbf{e}_{ij}, \qquad \bar{v}_j^{(g,ij)} = \bar{\mathbf{v}}_j^{(g)} \cdot \mathbf{e}_{ij}, \tag{6}$$

Message networks for advection and reaction are instantiated per channel group, so the learnt dynamics can specialise by physical regime. In the following, all barred quantities are per-group and, in the case of velocities, edge-projected.

Message mechanisms For a directed edge $(i \to j)$ with distance d_{ij}, the five mechanisms are

$$m_{\text{cont}} = \frac{\bar{\rho}_i \bar{v}_i}{d_{ij} + \epsilon}, \qquad m_{\text{diff}} = \frac{\bar{v}_i - \bar{v}_j}{d_{ij} + \epsilon}, \qquad m_{\text{relax}} = \frac{v_{\text{eq}}(\bar{\rho}_j) - \bar{v}_j}{\tau_j}, \tag{7}$$

$$m_{\text{advect}} = \frac{\eta_{ij}\, v_{ij}^{\text{up}}}{d_{ij} + \epsilon}, \qquad m_{\text{react}} = \Phi_{R,g}\big(\bar{\rho}_j,\, \bar{\rho}_j^{\text{trend}},\, \bar{q}_j\big). \tag{8}$$

The continuity message is the density contribution from the source node. Aggregated over a neighbourhood, the diffusion messages form a graph-Laplacian update with weights $(d_{ij} + \epsilon)^{-1}$. We use the inverse distance rather than the d_{ij}^{-2} scaling of a discrete second derivative. Relaxation drives the velocity towards the equilibrium $v_{\text{eq}}(\bar{\rho}_j) = \max(0,\, 1 - \bar{\rho}_j/2)$ on the timescale $\tau_j = \max\big(\epsilon, \exp(\theta_\tau + \beta\bar{\rho}_j)\big)$, with $\theta_\tau, \beta \in \mathbb{R}$ learnt and shared across nodes and channels. For traffic, this equilibrium has a direct Greenshields interpretation. For atmospheric groups, it acts as state-dependent damping. Advection is a learnt directional transport. Instead of discretising a velocity field, a group-specific network predicts an edge coefficient η_{ij} whose sign selects the upwind node through a differentiable gate:

$$\eta_{ij} = s_g\, \Phi_{\eta,g}\big(\bar{\rho}_i, \bar{\rho}_j, \bar{\rho}_i - \bar{\rho}_j, \bar{\rho}_i^{\text{trend}} - \bar{\rho}_j^{\text{trend}}, \bar{q}_i, \bar{q}_j\big), \qquad v_{ij}^{\text{up}} = \sigma(\kappa\eta_{ij})\, \bar{v}_i + \big(1 - \sigma(\kappa\eta_{ij})\big)\bar{v}_j. \tag{9}$$

The magnitude of η_{ij} sets the transport strength. κ sets the initial gate sharpness, and the learnt group scale $s_g = \exp(\theta_g) > 0$ adjusts both transport magnitude and effective sharpness. Because the message uses the upwind state rather than a state difference, transport between nodes with equal states remains nonzero, which also distinguishes advection from the difference-based diffusion. Reaction approximates the source $S(\rho)$ from destination-node statistics alone and is therefore constant across a node's incoming edges. For groups without a flux channel, the $\bar{q}$ arguments are omitted.

Physics-aware Edge Gating The mechanism-specific messages are stacked into $\mathbf{m}_{ij} = [m_{\text{cont}} \| m_{\text{visc}} \| m_{\text{relax}} \| m_{\text{advect}} \| m_{\text{react}}]$ and rescaled per mechanism by its standard deviation across edges, computed per sample and detached in the backward pass. Each normalised message is then gated independently, conditioned on the source and target states, edge distance, and normalised mechanism messages:

$$\boldsymbol{\gamma}_{ij} = \sigma\big(\mathrm{MLP}_{\gamma}\big([\mathbf{h}_i^{\text{avg}} \| \mathbf{h}_j^{\text{avg}} \| d_{ij} \| \hat{\mathbf{m}}_{ij}]\big)\big), \qquad \tilde{\mathbf{m}}_{ij} = \boldsymbol{\gamma}_{ij} \odot \hat{\mathbf{m}}_{ij}, \tag{10}$$

where $\mathbf{h}_i^{\text{avg}}$ averages $\mathbf{H}^{(l)}$ over channels and $\boldsymbol{\gamma}_{ij} \in [0,1]^5$ holds one gate per mechanism. The gates provide the adaptive, per-edge weighting that replaces a fixed operator sequence.

Node Aggregation and Channel-Aware Routing Gated messages are summed over incoming edges, giving one aggregate per mechanism and node, $\mathbf{M}_i = \sum_{j \in \mathcal{N}(i)} \tilde{\mathbf{m}}_{ji} \in \mathbb{R}^5$. A routing map assigns each mechanism to a set of hidden channels $\mathcal{C}_k \subseteq \{1, \ldots, C\}$, and the per-node physics update is scattered accordingly: channel c receives $\sum_{k:\, c \in \mathcal{C}_k} M_i^{(k)}$, and channels not targeted by any mechanism receive no physics update. Without a routing map the aggregate is broadcast to all channels. Routing therefore provides an explicit inductive bias along the physical-variable dimension: instead of unrestricted mixing between all channels, interactions are restricted according to their roles in the respective physical mechanisms.

Hidden-State Update The routed physics message is concatenated with the latent state and fused by an MLP,

$$\Delta\mathbf{h}_j = \mathrm{MLP}\big([\mathbf{h}_j \,\|\, \mathbf{m}_j]\big), \qquad \mathbf{h}'_j = \mathrm{LayerNorm}\big(\mathbf{h}_j + \Delta\mathbf{h}_j\big). \tag{11}$$

Both, parameter-free and learnt, mechanisms act only between channels assigned by $\mathcal{R}$, imposing physical structure on both messages and their connectivity.

4 EXPERIMENTS

To evaluate predictive performance and cross-domain generalizability, we consider highway traffic (PEMS04 and PEMS08 (California Department of Transportation (Caltrans), 2026)) and atmospheric systems (Beijing (Chen, 2017) and KnowAir (Wang et al., 2020)). The traffic datasets exhibit strong daily and weekly periodicity and spatial correlations along the road network. In contrast, the atmospheric datasets show annual seasonality, multi-day pollution episodes, and long-term concentration trends, with additional diurnal patterns in Beijing. Spatially, KnowAir spans 184 geographically dispersed stations with comparatively weak local correlations, whereas Beijing comprises 12 stations within a compact metropolitan region and exhibits stronger local dependencies. Since prior results did not cover a joint multivariate, multi-node forecasting setting, we adapt and retrain DLinear Zeng et al. (2023), STAEformer Liu et al. (2023), HG-GFNO Hosseini et al. (2026), AirDualODE Tian et al. (2025), DTRformer Chen et al. (2025), and AirDDE Wu et al. (2026) to jointly predict all channels and nodes (STD-MAE was not adapted as its two-stage univariate pretraining framework would require substantial unspecified methodological changes). Adaptations and baseline configurations are given in Appendix A.3 and Table 8. All models are evaluated under the same experimental setup. The temporal query uses 24-hour cycle lengths of 288, 24, and 8 steps for PEMS04/08, Beijing, and KnowAir, respectively. The input length T_{in} is selected per model and dataset on the validation set from $\{12, 24, 96\}$. On the traffic datasets, STAEformer, DTRformer, and AirDDE use an input window of $T_{\text{in}} = 12$, while CyPhy-STN, AirDualODE, HGGFNO, and DLinear use $T_{\text{in}} = 96$. On the atmospheric datasets, the two groups use $T_{\text{in}} = 24$ and $T_{\text{in}} = 96$, respectively (Appendix Table 5).We use 70/10/20 or 50/25/25 train/validation/test splits, and train for 120 epochs with AdamW, a common learning-rate schedule, and batch size 16. Full settings are reported in Appendix Table 8. CyPhy-STN uses the same architecture across domains, only assigning the domain-specific physical variables to the roles ρ, v, and q (Appendix Table 6). For traffic, these correspond to occupancy, speed, and flow, whereas for atmospheric data, pollutant and thermodynamic variables serve as transported scalars (ρ), while projected wind corresponds to the driving velocity (v).

Table 2: Component-level complementarity. Masked test MAE and equal-channel-weight σ-normalised MAE are reported as mean $\pm$ standard deviation over five seeds at the default forecast horizon. Bold/italic mark best/second best within each metric column.

Model	PEMS04		PEMS08		KnowAir		Beijing	
	MAE	σ-NMAE	MAE	σ-NMAE	MAE	σ-NMAE	MAE	σ-NMAE
External baselines								
DLinear	10.054 ± 0.026	0.213 ± 0.000	8.102 ± 0.005	0.199 ± 0.000	35.213 ± 0.222	0.613 ± 0.000	52.915 ± 0.044	0.648 ± 0.000
Air-DualODE Multiv.	7.747 ± 0.023	0.169 ± 0.001	5.900 ± 0.041	0.156 ± 0.001	44.122 ± 0.363	0.645 ± 0.001	53.313 ± 2.790	0.636 ± 0.003
HGGFNO	*6.780 ± 0.034*	*0.155 ± 0.001*	5.310 ± 0.041	0.149 ± 0.001	34.149 ± 0.515	0.513 ± 0.087	50.299 ± 0.338	*0.619 ± 0.001*
STAEformer	6.818 ± 0.031	*0.155 ± 0.000*	**5.212 ± 0.054**	**0.145 ± 0.001**	36.809 ± 1.631	**0.500 ± 0.102**	50.834 ± 0.375	0.635 ± 0.003
DTRformer	6.926 ± 0.100	*0.155 ± 0.001*	5.268 ± 0.015	*0.146 ± 0.000*	36.163 ± 1.573	0.597 ± 0.004	*50.149 ± 0.571*	0.624 ± 0.005
CyPhy-STN components								
CyPhy-STN	**6.735 ± 0.004**	**0.153 ± 0.000**	*5.233 ± 0.021*	*0.146 ± 0.000*	*32.474 ± 0.093*	*0.505 ± 0.090*	51.143 ± 0.594	0.621 ± 0.002
Temporal only	7.035 ± 0.017	0.161 ± 0.001	5.571 ± 0.017	0.152 ± 0.001	**32.127 ± 0.395**	0.506 ± 0.089	**49.561 ± 0.084**	**0.614 ± 0.001**
Physics GNN only	6.998 ± 0.016	0.158 ± 0.000	5.413 ± 0.032	0.149 ± 0.001	34.936 ± 0.525	0.519 ± 0.085	52.562 ± 0.176	0.635 ± 0.002

5 Evaluation

We evaluate CyPhy-STN along the main claims of our approach: consistent performance across different physical domains, complementary inductive biases from the structured decomposition of spatio-temporal dependencies, and the contribution of parallel, routed physics message passing. We first compare overall and cross-domain performance against state-of-the-art models. We then evaluate the individual contributions of the physics-inspired message passing mechanisms compared to long-horizon forecasting and analyze information propagation through the model. All experiments use five random seeds $24, 33, 900, 2028, 2029$, with results reported as mean $\pm$ standard deviation.

5.1 Overall and Cross-Domain Performance

Table 2 compares CyPhy-STN with recent state-of-the-art models on the joint prediction of all observed physical variables. We report masked test MAE and scale-normalised, equal-channel-weight σ-NMAE at the default forecast horizon. On the traffic datasets and KnowAir, CyPhy-STN ranks among the top two models in all six dataset–metric combinations and achieves the lowest error on PEMS04. On PEMS08, STAEformer achieves the lowest error with CyPhy-STN a close second. On KnowAir, differences among the leading models lie within one standard deviation across seeds. On Beijing, the temporal backbone of CyPhy-STN alone achieves the best result on both metrics, whereas the physics operators provide no additional benefit on this dataset (see Table 9). Overall, CyPhy-STN or one of its components ranks first in five of the eight dataset–metric combinations and within the top two in all of them, using a single architectural design across domains.

Table 3 reports errors for the primary target variables used in prior work (flow for PEMS04/08, PM2.5 for KnowAir/Beijing). In the multivariate setting, CyPhy-STN achieves the lowest target MAE and σ-NMAE on PEMS04 and is tied-best in σ-NMAE on PEMS08 (0.099, with STAEformer slightly lower in MAE, 14.37 vs. 14.42), with notably low seed variance on both traffic datasets. On KnowAir, the top multivariate models lie within 0.08 MAE of each other; the vanilla-attention variant is best (18.70), while the full model (18.79) is on par with STAEformer (18.76) and DTRformer (18.77). On Beijing, the variant without the reaction operator comes within 0.27 MAE of the best model, DTRformer, while the full model trails by 1.62 MAE. Specialised univariate models remain strongest on all three targets for which they are reported. Among univariate models, CyPhy-STN Univ. stays within 0.63 and 1.07 MAE of the best on PEMS04/08 but falls behind on KnowAir (+2.41 MAE), where AirDDE is strongest.

We next assess whether this performance is consistent across different domains rather than driven by strong results on individual datasets. Average rank captures a model's typical position among competing methods irrespective of performance margins. Following Chronos (Ansari et al., 2024), we aggregate error ratios via the geometric mean, but normalise by the best model per dataset instead of Seasonal Naive. For each dataset d, let $E_d^\star = \min_{j \in \mathcal{M}} E_{j,d}$ denote the lowest (seed-averaged) error among all compared models $\mathcal{M}$. We define the geometric relative regret and, since the geometric mean can mask a single poor result, the worst-case relative regret of model m over D datasets as

$$\mathrm{GRR}\, m = 100 \left[\left(\prod d = 1^D \frac{E_{m,d}}{E_d^\star} \right)^{1/D} - 1 \right] \qquad \mathrm{RR}^{\max}\, m = 100 \max d \left(\frac{E_{m,d}}{E_d^\star} - 1 \right) \quad (12)$$

Table 3: Target-variable test MAE only. PEMS04/08 report Flow, KnowAir/Beijing report PM2.5. Consistent with previous papers values are averaged over the default forecast horizon. **Best** and *second best* are marked separately within the uni-variate and multi-variate groups. Cy-Phy vanilla uses full softmax attention instead of a cyclic query.

Model	PEMS04 (Flow (q))		PEMS08 (Flow (q))		KnowAir (PM2.5)		Beijing (PM2.5)	
	MAE	σ-NMAE	MAE	σ-NMAE	MAE	σ-NMAE	MAE	σ-NMAE
Multivariate models								
DLinear Multiv.	28.23 ± 0.08	0.181 ± 0.001	22.72 ± 0.02	0.156 ± 0.000	19.69 ± 0.00	0.408 ± 0.000	41.87 ± 0.06	0.526 ± 0.001
Air-DualODE Multiv.	21.66 ± 0.06	0.139 ± 0.000	16.38 ± 0.11	0.113 ± 0.001	19.20 ± 0.12	0.398 ± 0.002	38.36 ± 0.68	0.482 ± 0.009
HGGFNO Multiv.	18.81 ± 0.10	**0.120 ± 0.001**	14.61 ± 0.13	*0.100 ± 0.001*	18.96 ± 0.08	0.393 ± 0.002	37.99 ± 0.37	0.478 ± 0.005
STAE Multiv.	18.95 ± 0.09	*0.121 ± 0.001*	**14.37 ± 0.16**	**0.099 ± 0.001**	18.76 ± 0.17	0.389 ± 0.004	*37.68 ± 0.45*	*0.474 ± 0.006*
DTRformer Multiv.	19.28 ± 0.30	0.123 ± 0.002	14.54 ± 0.05	*0.100 ± 0.000*	18.77 ± 0.28	0.389 ± 0.006	**37.48 ± 0.29**	**0.471 ± 0.004**
AirDDE Multiv.	23.51 ± 1.25	0.150 ± 0.008	18.30 ± 0.40	0.126 ± 0.003	19.14 ± 0.22	0.396 ± 0.005	40.45 ± 0.73	0.509 ± 0.009
CyPhy-STN (ours)	*18.70 ± 0.01*	**0.120 ± 0.000**	*14.42 ± 0.06*	**0.099 ± 0.000**	18.75 ± 0.09	*0.388 ± 0.002*	39.17 ± 0.75	0.493 ± 0.009
CyPhy-STN vanilla	18.88 ± 0.02	*0.121 ± 0.000*	14.45 ± 0.06	**0.099 ± 0.000**	**18.70 ± 0.11**	**0.387 ± 0.002**	39.04 ± 0.39	0.491 ± 0.005
CyPhy-STN w/o cont.	18.73 ± 0.06	**0.120 ± 0.000**	14.45 ± 0.02	**0.099 ± 0.000**	18.76 ± 0.07	*0.388 ± 0.001*	39.22 ± 0.65	0.493 ± 0.008
CyPhy-STN w/o visc.	*18.70 ± 0.01*	**0.120 ± 0.000**	14.43 ± 0.05	**0.099 ± 0.000**	18.78 ± 0.13	0.389 ± 0.003	39.25 ± 0.70	0.494 ± 0.009
CyPhy-STN w/o relax.	**18.69 ± 0.02**	**0.120 ± 0.000**	14.43 ± 0.08	**0.099 ± 0.001**	18.74 ± 0.07	*0.388 ± 0.001*	38.53 ± 0.83	0.485 ± 0.010
CyPhy-STN w/o adv.	18.71 ± 0.02	**0.120 ± 0.000**	14.46 ± 0.11	**0.099 ± 0.001**	18.79 ± 0.11	0.389 ± 0.002	38.98 ± 0.67	0.490 ± 0.008
CyPhy-STN w/o react.	18.85 ± 0.03	*0.121 ± 0.000*	14.57 ± 0.11	*0.100 ± 0.001*	*18.73 ± 0.07*	*0.388 ± 0.001*	37.82 ± 0.51	0.476 ± 0.006
CyPhy-STN w/o temp.	19.46 ± 0.05	0.125 ± 0.000	14.95 ± 0.09	0.103 ± 0.001	19.30 ± 0.02	0.400 ± 0.000	40.96 ± 0.98	0.515 ± 0.012
Temporal	19.54 ± 0.05	0.125 ± 0.000	15.41 ± 0.04	0.106 ± 0.000	18.84 ± 0.03	0.390 ± 0.001	38.26 ± 0.22	0.481 ± 0.003
Univariate models								
FEDformer	29.30	–	25.00	–	–	–	–	–
iTransformer	23.30	–	18.00	–	–	–	–	–
STAEformer Univ.	18.22	–	13.46	–	–	–	–	–
STD-MAE Univ.	**17.80**	–	*13.44*	–	–	–	–	–
DTRformer Univ.	*18.00*	–	**13.17**	–	–	–	–	–
AirDualODE	–	–	–	–	*18.64*	–	–	–
AirDDE Univ.	–	–	–	–	**16.92**	–	–	–
CyPhy-STN Univ.	18.43	0.118	14.24	0.098	19.33	0.400	–	–

A GRR_m of zero indicates that model m is best on every dataset. The cross-domain results are summarised in Table 4. CyPhy-STN achieves the lowest average rank (2.00), mean relative regret (0.72%), and worst-case relative regret (2.12%) in the seven-model comparison. Across the evaluated traffic and atmospheric datasets, its error stays within 2.12% of the best model on every dataset. STAEformer ranks second on average rank and mean regret (2.25 and 1.15%), followed by HGGFNO (3.00 and 1.83%). The metrics capture complementary aspects of cross-domain performance, e.g., AirDDE has a worse average rank than DTRformer (5.00 vs. 3.75), a higher mean regret (10.07% vs. 5.76%), but a lower worst-case regret (17.56% vs. 20.57%), as DTRformer degrades strongly on KnowAir. Figure 2 relates this cross-domain consistency to model size. CyPhy-STN attains a best average rank while using fewer parameters than STAEformer, DTRformer, and HGGFNO. DLinear is considerably smaller but ranks last. These results show a cross-domain advantage of CyPhy-STN, as it remains consistently competitive across systems with different physical dynamics, without the need for a large parameter budget or domain-specific architectural redesign.

Table 4: Cross-domain summary of masked test σ-NMAE. Average rank is taken over datasets. RR is the relative gap (%) to the best model per dataset, with mean and worst RR over datasets.

Model	Avg Rank ↓	Mean RR ↓	Worst RR ↓
CyPhy-STN	**2.00**	**0.72**	**2.12**
STAEformer	*2.25*	*1.15*	*2.71*
HGGFNO	3.00	1.83	3.71
DTRformer	3.75	5.76	20.57
AirDDE	5.00	10.07	17.56
Air-DualODE Multiv.	5.50	12.59	30.26
DLinear	6.50	26.08	38.57

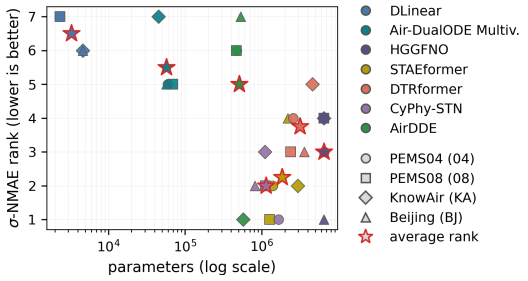


Figure 2: Parameter efficiency across domains: σ-NMAE rank per dataset and average rank versus parameter count.

5.2 Component and Physics Mechanism Ablation Studies

We next examine whether separating temporal and physics-based pathways provides complementary inductive biases (Table 2). The two pathways contribute differently across domains. On traffic, the physics-GNN-only variant outperforms the temporal-only variant (TC-MHA) on both PEMS04

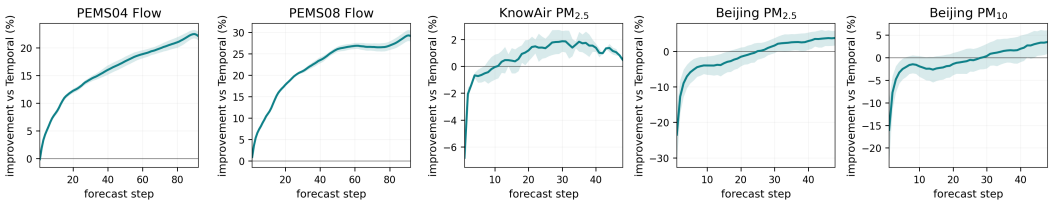


Figure 3: Relative MAE improvement of CyPhy-STN over the temporal-only variant, $(\mathrm{MAE}_{\mathrm{Temporal}} - \mathrm{MAE}_{\mathrm{CyPhy}})/\mathrm{MAE}_{\mathrm{Temporal}} \times 100$, positive favours CyPhy-STN. Horizons: 92 steps (PEMS04/08), 48 (KnowAir/Beijing). Shading: seed std.

and PEMS08. Combining both yields the best results and stays within 4% of the specialised traffic models. On the atmospheric datasets, the order reverses. The temporal-only variant outperforms the physics-GNN-only variant and even slightly outperforms the full model in KnowAir MAE and Beijing σ-NMAE. Graph-based physical interactions thus contribute most on traffic, whereas temporal dependencies dominate on the atmospheric datasets. The operator-level ablations in Appendix A.2.1 are consistent with this. Physics-informed message passing improves traffic forecasting, whereas on the atmospheric datasets the effects of individual operators mostly lie within seed variation. The target-variable ablations (Table 3) confirm the central role of the temporal component. Removing it causes the largest and most consistent degradation, increasing MAE by 0.5 to 1.75 on all datasets. Removing individual physics terms changes MAE only marginally. Reaction has the clearest effect on traffic, relaxation on Beijing. Overall, temporal dependencies and routed physics-based interactions make distinct contributions, whose importance depends on the physical domain.

5.3 Long-Horizon Forecasting and Information Propagation

Although physics-informed graph models improve roll-out stability, the impact of message passing across varying forecast horizons remains unclear. In CyPhy-STN, temporal dependencies and physics-based spatial and cross-variable interactions are modelled through distinct pathways. The benefit of these may grow with the forecast horizon, as long-range predictions increasingly depend on propagated interactions that the temporal connection alone cannot capture. To test this, we extend the forecast horizon, and compare the full CyPhy-STN with its no-physics variant (Figure 3). On traffic, the gain grows almost steadily with the horizon. It reaches about 22% on PEMS04 and 29% on PEMS08 at the final step. On the atmospheric datasets, physics-informed message passing degrades the first forecast steps but becomes beneficial from around step 10 on KnowAir and step 25–30 on Beijing. KnowAir peaks at about 2% near step 30 before declining again. The benefit generally increases with the horizon, although not monotonically across all datasets.
To understand how information propagates through the distinct pathways, we analyse three complementary Jacobians: *end-to-end* ($\mathbf{X} \to \widehat{\mathbf{Y}}$), *hidden-state* ($\mathbf{H}^{(0)} \to \mathbf{H}^{(L)}$), and *physics-input* ($\mathbf{X}_{\mathrm{phys}} \to \mathbf{H}^{(L)}$). These reveal strong local propagation within the physics-message stack, while the complete model retains substantial long-range sensitivity. For detailed analysis, see Appendix A.2.2.

6 Conclusion

We introduced CyPhy-STN, a unified architecture for multivariate spatio-temporal forecasting that combines physics-inspired message passing with cyclic temporal attention. Rather than treating graph propagation as a generic aggregation process or prescribing a fixed governing equation, CyPhy-STN decomposes spatial interactions into interpretable physical mechanisms and combines them through channel-aware routing and adaptive edge-wise gating. Across traffic and atmospheric forecasting, CyPhy-STN demonstrates competitive performance without domain-specific architectural modifications, attaining the best cross-domain average rank and relative regret among evaluated models. Our ablations show that temporal and physics-based pathways provide complementary but domain-dependent inductive biases. Our analysis of information-propagation shows strong local physical coupling while the model retains sensitivity across larger graph distances. Together, these results indicate that separating temporal modelling from routed physical interactions provides a flexible framework for forecasting across systems with different spatio-temporal dynamics.

AI USE STATEMENT

In this work, generative AI tools were used to provide secondary feedback on independently developed experimental designs and interpretations of results, and to assist with language editing and Latex formatting. Generative AI was not used to originate research questions, model architecture, or methodological contributions, nor to generate synthetic data or prove mathematical claims. All AI-assisted suggestions were critically reviewed by the authors, and experimental decisions, analyses, and conclusions were independently verified. The authors take full responsibility for the final content of the paper, including all text, claims, and artefacts produced with the help of generative AI.

REPRODUCIBILITY STATEMENT

The complete source code, including model implementations, SotA model adaptations, training and evaluation scripts, the raw data and the data loading and preprocessing pipeline, is provided as anonymised supplementary material and will be released publicly upon acceptance. We used fixed random seeds for all experiments and report the hyper-parameters, and experimental settings in the Appendix. They can also be found in the configuration files in the supplementary material. We report results as mean and standard deviation over 5 runs. The software environment is managed with `uv`. All package dependencies and their versions are specified in the `pyproject.toml`, and `requirements.txt` files included in the repository. The environment can be recreated directly from these files.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Nikolas Assiotis, Rachel Hau, Valentijn Oldenburg, Rik Verbiest, Julian Koellermeier, Matthia Sabatelli, and Juan Cardenas-Cartagena. Physics-informed graph neural networks for air pollution forecasting in the Netherlands. In Cecília Coelho, Bernd Zimmering, M. Fernanda P. Costa, Luís L. Ferrás, and Oliver Niggemann (eds.), *Proceedings of the 2nd ECAI Workshop on "Machine Learning Meets Differential Equations: From Theory to Applications"*, volume 277 of *Proceedings of Machine Learning Research*, pp. 47–70. PMLR, 2025. URL https://proceedings.mlr.press/v277/assiotis25a.html.

California Department of Transportation (Caltrans). PeMS04 and PeMS08 traffic flow datasets for traffic flow forecasting. IEEE DataPort, 2026. URL https://doi.org/10.21227/m4p8-ag97.

Benjamin Paul Chamberlain, James Rowbottom, Maria I. Gorinova, Stefan D. Webb, Emanuele Rossi, and Michael M. Bronstein. GRAND: Graph neural diffusion. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1407–1418. PMLR, 2021. URL https://proceedings.mlr.press/v139/chamberlain21a.html.

Jing Chen, Haocheng Ye, Zhian Ying, Yuntao Sun, and Wenqiang Xu. Dynamic trend fusion module for traffic flow prediction. *Applied Soft Computing*, 174:112979, 2025. doi: 10.1016/j.asoc.2025.112979. URL https://doi.org/10.1016/j.asoc.2025.112979.

Songxi Chen. Beijing Multi-Site Air Quality. UCI Machine Learning Repository, 2017. URL https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data.

Moshe Eliasof, Eldad Haber, and Eran Treister. Feature transportation improves graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11874–

11882, 2024. doi: 10.1609/aaai.v38i11.29073. URL https://doi.org/10.1609/aaai.v38i11.29073.

Haotian Gao, Renhe Jiang, Zheng Dong, Jinliang Deng, Yuxin Ma, and Xuan Song. Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3998–4006, 2024.

Seyed-Majid Hosseini, S. Mozhgan Rahmatinia, and Seyed-Amin Hosseini-Seno. Integrated spatio-temporal modeling with hybrid graph convolutions and the graph Fourier neural operator for traffic prediction. *Scientific Reports*, 16(1):12945, 2026. doi: 10.1038/s41598-026-38563-y. URL https://doi.org/10.1038/s41598-026-38563-y.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866. URL https://doi.org/10.1080/03610919008812866.

Zheng Jiang, Wei Wang, Gaowei Zhang, and Yi Wang. SSTODE: Ocean-atmosphere physics-informed neural ODEs for sea surface temperature prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38661–38669, 2026. doi: 10.1609/aaai.v40i45.41209. URL https://doi.org/10.1609/aaai.v40i45.41209.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=e24CueVty2.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla Transformer SOTA for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pp. 4125–4129, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3583780.3615160. URL https://doi.org/10.1145/3583780.3615160.

Rongye Shi, Zhaobin Mo, and Xuan Di. Physics-informed deep learning for traffic state estimation: A hybrid paradigm informed by second-order traffic models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 540–547, 2021. doi: 10.1609/aaai.v35i1.16132. URL https://doi.org/10.1609/aaai.v35i1.16132.

Jindong Tian, Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, and Bin Yang. Air quality prediction with physics-guided dual neural ODEs in open systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kOJf7Dklyv.

Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. PM2.5-GNN: A domain knowledge enhanced graph neural network for PM2.5 forecasting. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '20, pp. 163–166, 2020. doi: 10.1145/3397536.3422208. URL https://doi.org/10.1145/3397536.3422208.

Zhihong Wang, Wei Li, Zhuoxuan Liang, Junhui Jiang, Xiaohua Jia, and Moustafa Youssef. MSTHH: A unified framework for asynchronous and heterogeneous multimodal traffic prediction. *Information Sciences*, 754:123652, 2026. doi: 10.1016/j.ins.2026.123652. URL https://doi.org/10.1016/j.ins.2026.123652.

Binqing Wu, Zongjiang Shang, Shiyu Liu, Jianlong Huang, Jiahui Xu, and Ling Chen. AirDDE: Multifactor neural delay differential equations for air quality forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15941–15949, 2026. doi: 10.1609/aaai.v40i19.38627. URL https://doi.org/10.1609/aaai.v40i19.38627.

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*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Table 5: Validation-selected input lengths T_{in} used for each model–dataset pair.

Model	PEMS04	PEMS08	KnowAir	Beijing
CyPhy-STN	96	96	96	96
STAEformer	12	12	24	24
DTRformer	12	12	24	24
AirDDE	12	12	24	24
AirDual	96	96	96	96
DLinear	96	96	96	96
HGGFNO	96	96	96	96

A Appendix

A.1 Detailed training settings

Input length. The baselines differ substantially in the lookback windows for which they are designed, such that enforcing a single T_{in} across all models could systematically disadvantage architectures tuned for either short- or long-context forecasting. We therefore select T_{in} independently for each model–dataset pair from $12, 24, 96$ using validation MAE and report test performance only for the validation-selected configuration. The resulting input lengths are listed in Table 5. STAEformer, DTRformer, and AirDDE consistently select shorter input windows. CyPhy-STN, DLinear, AirDual, and HGGFNO select a 96-step window across all datasets. This procedure allows each architecture to operate at its preferred temporal context without using test performance for hyperparameter selection.

Physics configuration. Table 6 summarises the channel roles, regime grouping, and output routing used by CyPhy-STN. The datasets have substantial differences regarding their physical channels. For PEMS04 and PEMS08, the three observed variables map directly to the transported-scalar (ρ), velocity (v), and flux (q) roles, and all mechanism outputs are broadcast to the latent state. KnowAir contains a heterogeneous set of meteorological variables. Twelve channels share the transported-scalar role, while the wind variables form two velocity regimes corresponding to different atmospheric levels. This allows distinct regime-specific dynamics to be represented within the same graph. Beijing uses the most selective routing scheme: the six pollutant channels receive the physics updates, while meteorological variables act as physical drivers of the messages but are not themselves updated by the physics branch. Channels without an assigned physical role remain exclusively on the temporal path.

Graph statistics. The graph structures used by the four datasets are summarised in Table 7. They span different spatial regimes, ranging from the small and comparatively dense Beijing graph ($N = 12$, density 0.3636) to the substantially larger and sparse PEMS04 graph ($N = 307$, density 0.0036). Their shortest-path structure also differs considerably: Beijing has a maximum finite directed shortest-path distance of four, whereas the diameter of PEMS04 reaches 53. The reachable-pair counts at each shortest-path distance quantify the number of node pairs available in each distance shell of the Jacobian sensitivity analysis. In particular, Beijing contains no reachable pair beyond $d = 4$, whereas the larger graphs support substantially longer-range sensitivity measurements.

Optimization settings. The remaining training and optimization hyperparameters are summarised in Table 8. Unless stated otherwise, these settings are shared across datasets. Dataset-specific differences are restricted to dropout, the mechanism parameter δ, the number of samples used for the Jacobian estimator, and the cycle length P. The latter corresponds to a 24-hour periodicity under the sampling frequency of each dataset: $P = 288$ for PEMS04/08, $P = 8$ for KnowAir, and $P = 24$ for Beijing.

Table 6: Channel-aware physics configuration. Each input channel is assigned a physical role (ρ transported scalar, v velocity, q flux) that determines how it enters message computation. Channels are partitioned into regime groups, each with its own velocity field and its own advection and reaction networks. Mechanism outputs are either broadcast across all latent channels or routed to selected targets, channels with no role remain on the temporal path.

Dataset	Role	Channels	Target
PEMS04/08	ρ	occupancy	broadcast
	v	speed	broadcast
	q	flow	broadcast
KnowAir	ρ	PM2.5, 2m temperature, 2m dewpoint, boundary-layer height, K index, RH 950/975hPa, SH 950hPa, surface pressure, temperature 925/950hPa, total precipitation	broadcast
	v (surface)	100m u-wind, 100m v-wind	broadcast
	v (950hPa)	u-wind 950hPa, v-wind 950hPa	broadcast
	q	vertical velocity 950hPa, vorticity 950hPa	broadcast
Beijing	ρ	PM2.5, PM10, SO_2, NO_2, CO, O_3	channels 0–5
	v	wind u, wind v	channels 0–5
	—	temperature, pressure, dewpoint, rain, wind speed	temporal path

Table 7: Dataset configurations and graph statistics. Density is $|\mathcal{E}|/[N(N-1)]$, D_{SP} and $\bar{d}_{\mathrm{SP}}$ denote the maximum and mean finite directed shortest-path distances. The right-hand block reports the number of reachable directed node pairs at each shortest-path distance and therefore bounds the shell populations used in the Jacobian analysis.

								Reachable pairs at d				
Dataset	Domain	N	$\lvert\mathcal{E}\rvert$	Dens.	D_{SP}	$\bar{d}_{\mathrm{SP}}$	C	1	2	3	4	5
PEMS04	Traffic	307	340	0.0036	53	15.56	3	340	375	406	431	464
PEMS08	Traffic	170	277	0.0096	25	8.41	3	277	425	629	866	1124
Beijing	Atmosphere	12	48	0.3636	4	1.69	13	48	37	13	2	0
KnowAir	Atmosphere	184	1472	0.0437	16	5.72	18	1472	2689	3556	4141	4488

A.2 Detailed Experimental Results

A.2.1 Ablation studies

Contribution of physics-informed message passing. The leave-one-out ablation in Table 9 shows distinct, domain-dependent contributions of the physics operators. In both traffic datasets, the removal of all physics operators consistently degrades the prediction performance, with increases in σ-NMAE of 0.0073 and 0.0061 in PEMS04 and PEMS08, respectively. These degradations exceed both the observed seed variability and the differences in the rerun of the same-seed, providing evidence that physics-informed message passing contributes beyond temporal modelling alone. Among individual operators, reaction exhibits the most consistent contribution to traffic forecasting, whereas the remaining operators show smaller or less consistent effects. In contrast, the atmospheric datasets exhibit a different pattern. On KnowAir, removing individual operators produces changes that are generally small relative to seed variability, while on Beijing, several operator removals improve forecasting performance. These results indicate that the benefits of physics-informed message passing are neither uniform across operators nor universally transferable across domains. Rather, their contribution depends on the interaction between the imposed physical inductive biases and the underlying system dynamics.

A.2.2 Information Preservation and Graph Bottlenecks

To analyse information propagation and potential over-squashing, we estimate distance-resolved Jacobian Frobenius norms using Hutchinson's stochastic estimator with Rademacher probes Hutchin-

Table 8: Training and optimization hyperparameters. Values are shared across datasets except where noted, per-dataset overrides are listed below the rule.

Hyperparameter	Value
Max epochs	120
Batch size B	16
Optimizer	AdamW
Initial / minimum learning rate	2×10^{-4} / 1×10^{-6}
Warmup / hold epochs	10 / 5
LR scheduler	sequential cosine
Early stopping patience / Δ	30 / 10^{-4}
Loss	masked MAE
Feature scaler	standard scaling
Gradient clipping	2.0
Per-dataset	
Dropout	0.1 (PEMS08), 0.2 (otherwise)
δ	0.5 (PEMS08), 1.0 (otherwise)
Jacobian samples	1 (PEMS08), 8 (otherwise)
Cycle length P	288 (PEMS), 8 (KnowAir), 24 (Beijing)

Table 9: Paired physics-operator leave-one-out ablation deltas relative to the full CyPhy-STN model. Each dataset reports equal-channel-weight masked test $\Delta\sigma$-NMAE and ΔMASE as mean $\pm$ standard deviation over matched seeds (24, 33, 900, 2028, 2029). Positive values (red) indicate degradation upon removal, while negative values (green) indicate improvement. Darker shading highlights more pronounced effects. † denotes a one-sided paired t-test with $p < 0.05$ for degradation.

	PEMS04		PEMS08		KnowAir		Beijing	
Variant	$\Delta\sigma$-NMAE	ΔMASE	$\Delta\sigma$-NMAE	ΔMASE	$\Delta\sigma$-NMAE	ΔMASE	$\Delta\sigma$-NMAE	ΔMASE
Same-seed rerun $\lvert\Delta\rvert$	0.0000	0.0001	0.0001	0.0000	0.0001	0.0017	0.0001	0.0002
w/o advection	-0.0001 $\pm$ 0.0001	-0.0001 $\pm$ 0.0004	+0.0001 $\pm$ 0.0006	+0.0004 $\pm$ 0.0022	-0.0022 $\pm$ 0.0030	-0.0049 $\pm$ 0.0068	-0.0007 $\pm$ 0.0006	-0.0011 $\pm$ 0.0010
w/o continuity	+0.0000 $\pm$ 0.0001	+0.0001 $\pm$ 0.0005	-0.0004 $\pm$ 0.0005	-0.0008 $\pm$ 0.0017	+0.0003 $\pm$ 0.0025	-0.0004 $\pm$ 0.0044	+0.0008 $\pm$ 0.0018	-0.0002 $\pm$ 0.0014
w/o relaxation	+0.0000 $\pm$ 0.0002	-0.0001 $\pm$ 0.0008	-0.0002 $\pm$ 0.0007	-0.0004 $\pm$ 0.0023	+0.0003 $\pm$ 0.0026	+0.0028 $\pm$ 0.0077	-0.0021 $\pm$ 0.0025	-0.0011 $\pm$ 0.0083
w/o reaction	+0.0013 $\pm$ 0.0003†	+0.0040 $\pm$ 0.0009†	+0.0013 $\pm$ 0.0008†	+0.0041 $\pm$ 0.0023†	-0.0029 $\pm$ 0.0046	-0.0051 $\pm$ 0.0102	-0.0057 $\pm$ 0.0023	-0.0105 $\pm$ 0.0069
w/o viscosity	+0.0001 $\pm$ 0.0001†	+0.0003 $\pm$ 0.0002†	-0.0001 $\pm$ 0.0008	-0.0002 $\pm$ 0.0024	-0.0009 $\pm$ 0.0023	-0.0014 $\pm$ 0.0051	+0.0006 $\pm$ 0.0008	+0.0003 $\pm$ 0.0005
w/o all physics	+0.0073 $\pm$ 0.0006†	+0.0218 $\pm$ 0.0018†	+0.0061 $\pm$ 0.0008†	+0.0201 $\pm$ 0.0025†	+0.0004 $\pm$ 0.0036	-0.0001 $\pm$ 0.0081	-0.0065 $\pm$ 0.0023	-0.0135 $\pm$ 0.0060

son (1990). For a Rademacher vector $\mathbf{v}$ drawn over the output entries of a target node j, $\mathbb{E}_{\mathbf{v}}\left\|\mathbf{v}^\top \partial \mathbf{Z}_j^{(L)} / \partial \mathbf{Z}_i^{(0)}\right\|_2^2 = \left\|\partial \mathbf{Z}_j^{(L)} / \partial \mathbf{Z}_i^{(0)}\right\|_F^2$. A single vector–Jacobian product per probe and target therefore yields unbiased sensitivity estimates for all source nodes at once. Source nodes are grouped by their directed shortest-path distance $d(i, j)$ to the target. Unreachable pairs are excluded. Writing $\mathbf{Z}^{(0)}$ and $\mathbf{Z}^{(L)}$ for the source and target representations of a given Jacobian variant, and indexing source and target nodes by i and j, we define

$$\mathcal{J}(d) = \left(\frac{1}{|\mathcal{P}_d|} \sum_{(i,j)\in\mathcal{P}_d} \left\| \frac{\partial \mathbf{Z}_j^{(L)}}{\partial \mathbf{Z}_i^{(0)}} \right\|_F^2 \right)^{1/2}, \tag{13}$$

$$\mathcal{P}_d = \left\{ (i,j) \in \mathcal{V} \times \mathcal{T} \;\middle|\; d(i,j) = d \right\}, \tag{14}$$

where $\mathcal{T} \subseteq \mathcal{V}$ is a fixed set of sampled target nodes, reused across all probes. Squared sensitivities are pooled within each distance shell before taking the square root. Absolute sensitivity scales differ across Jacobian variants and models. We therefore analyse the normalised profile

$$\rho(d) = \frac{\mathcal{J}(d)}{\mathcal{J}(1)}. \tag{15}$$

To quantify decay beyond the first hop, we fit

$$\ln \rho(d) = \alpha + \beta\, d, \qquad 2 \le d \le L - 1, \tag{16}$$

by least squares with a free intercept α. Excluding $d = 1$ restricts the fit to distances at which influence must traverse composed message passing. Excluding $d = L$ removes the boundary of the receptive field of the $L = 5$ one-hop physics-message layers. At this boundary, influence can only arrive through paths that perform a message step in every layer. Since α absorbs the normalisation,

Table 10: Distance-resolved Jacobian sensitivity and Dirichlet smoothing of CyPhy-STN. $\rho(2) = \mathcal{J}(2)/\mathcal{J}(1)$ quantifies the first-hop drop; β is the slope of $\ln \rho(d)$ over $2 \le d \le 4$ (Eq. equation 16), with per-hop retention e^{β}; $\mathcal{D}$ is the after/before ratio of normalised Dirichlet energy, averaged over physics-message layers. Mean $\pm$ std over 3 seeds.

	End-to-end		Hidden-state		Physics-input		
Dataset	$\rho(2)$	β	$\rho(2)$	β	$\rho(2)$	β	$\mathcal{D}$
PEMS04	0.097 ± 0.035	-0.56 ± 0.21	0.071 ± 0.009	-3.45 ± 0.30	0.062 ± 0.004	-3.42 ± 0.28	0.80 ± 0.01
PEMS08	0.048 ± 0.022	-0.66 ± 0.13	0.060 ± 0.024	-4.16 ± 0.14	0.030 ± 0.010	-4.03 ± 0.12	0.85 ± 0.00
KnowAir	0.365 ± 0.069	0.11 ± 0.07	0.111 ± 0.031	-2.98 ± 0.49	0.093 ± 0.019	-3.00 ± 0.47	0.91 ± 0.02
Beijing	0.871 ± 0.032	-0.07 ± 0.14	0.299 ± 0.032	-1.40 ± 0.43	0.371 ± 0.046	-1.32 ± 0.62	0.96 ± 0.00

β does not depend on the reference distance. We report $\rho(2)$ for the first-hop drop and β with its per-hop retention e^{β}. Means and standard deviations are computed across three seeds. We use $P = 16$ probes, $|\mathcal{T}| = 16$ target nodes, and distance shells up to $d = 5$. A strongly negative β ($e^{\beta} \ll 1$) indicates that the influence of distant nodes vanishes geometrically, which is consistent with over-squashing. A value of $\beta \approx 0$ indicates that distant nodes retain an influence comparable to that of nearby ones.

We evaluate three complementary Jacobian variants:

- *end-to-end*, from the input sequence $\mathbf{X}$ to the prediction $\widehat{\mathbf{Y}}$, capturing all propagation paths,
- *hidden-state*, from the input of the first physics-message layer $\mathbf{H}^{(0)}$ to the output of the final layer $\mathbf{H}^{(L)}$, capturing composed message propagation,
- *physics-input*, from the physical input $\mathbf{X}_{\text{phys}}$ to $\mathbf{H}^{(L)}$, additionally capturing the repeated injection of physical information into the physics-message stack.

For CyPhy-STN, weak multi-hop hidden-state sensitivity does not imply weak physical message passing. The physics-message layers are constructed as local, discretisation-inspired operators, in which neighbouring physical states determine mechanism-specific updates through continuity, advection, diffusion, relaxation, and reaction terms. Thus, strong one-hop sensitivity reflects effective local coupling of physically structured information.

To quantify over-smoothing, we additionally compute the Frobenius-normalised Dirichlet energy of a hidden representation $\mathbf{H}$,

$$\mathcal{D}_{\text{norm}}(\mathbf{H}) = \frac{\dfrac{1}{|\mathcal{E}|} \displaystyle\sum_{(i,j)\in\mathcal{E}} \dfrac{\left\|\mathbf{H}_i - \mathbf{H}_j\right\|_F^2}{\ell_{ij}}}{\dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \left\|\mathbf{H}_i\right\|_F^2 + \varepsilon}, \tag{17}$$

where ℓ_{ij} is the physical length of edge (i, j) and ε is a small constant for numerical stability. Normalising by the mean squared node norm makes equation 17 approximately invariant to the overall scale of $\mathbf{H}$. Values thus remain comparable across layers and models. We report the ratio $\mathcal{D}$ of the Dirichlet energy after each physics-message layer to that before it, averaged over layers. Ratios below one indicate reduced variation between neighbouring representations and thus smoothing.

Table 10 summarises the results for CyPhy-STN. Within the physics-message stack, sensitivity is strongly local. On PEMS04, PEMS08, and KnowAir, the hidden-state Jacobian drops to 6–11% of its one-hop value at $d = 2$. Beyond that, β lies between -4.16 and -2.98, i.e. only 1.6–5.1% of the sensitivity is retained per hop. The physics-input Jacobian follows almost the same profile, with β between -4.03 and -3.00. Its $\rho(2)$ is even lower on three of the four datasets. Although $\mathbf{X}_{\text{phys}}$ enters every physics-message layer, this repeated injection therefore does not extend the spatial reach of physical information.

Beijing shows weaker attenuation, with $\rho(2) = 0.30$ and a retention of 25% per hop for the hidden state. However, its graph is substantially smaller and denser (see Appendix, Table 7). Its outermost

Table 11: Decay slope β of all Jacobian variants and Dirichlet ratio $\mathcal{D}$ across ablations (mean $\pm$ std over seeds).

Dataset	Variant	End-to-end	Hidden-state	Physics-input	$\mathcal{D}$
PEMS04	CyPhy-STN	-0.56 ±0.21	-3.45 ±0.30	-3.42 ±0.28	0.80 ±0.01
	vanilla attn.	-0.76 ±0.16	-3.64 ±0.19	-3.79 ±0.48	0.81 ±0.00
	w/o continuity	-0.82 ±0.24	-3.34 ±0.30	-3.22 ±0.15	0.79 ±0.01
	w/o diffusion	-0.58 ±0.15	-3.29 ±0.27	-3.55 ±0.30	0.79 ±0.01
	w/o relaxation	-0.47 ±0.06	-3.35 ±0.52	-3.47 ±0.49	0.79 ±0.01
	w/o advection	-0.56 ±0.09	-3.49 ±0.30	-3.74 ±0.32	0.80 ±0.01
	w/o reaction	-0.28 ±0.26	-4.19 ±0.48	-3.82 ±1.03	0.80 ±0.02
	w/o temporal	-3.06 ±0.77	-3.30 ±0.41	-3.44 ±0.73	0.81 ±0.01
PEMS08	CyPhy-STN	-0.66 ±0.13	-4.16 ±0.14	-4.03 ±0.12	0.85 ±0.00
	vanilla attn.	-0.66 ±0.07	-4.18 ±0.75	-4.36 ±0.62	0.87 ±0.01
	w/o continuity	-0.75 ±0.21	-4.64 ±0.26	-4.63 ±0.26	0.84 ±0.01
	w/o diffusion	-0.64 ±0.11	-4.28 ±0.39	-3.98 ±0.15	0.84 ±0.00
	w/o relaxation	-0.66 ±0.57	-4.41 ±0.57	-4.68 ±0.59	0.85 ±0.00
	w/o advection	-0.61 ±0.16	-4.48 ±0.36	-4.68 ±0.26	0.84 ±0.00
	w/o reaction	-0.03 ±0.02	-4.76 ±0.24	-4.58 ±0.53	0.84 ±0.01
	w/o temporal	-4.01 ±0.03	-3.72 ±0.24	-3.98 ±0.54	0.85 ±0.01
KnowAir	CyPhy-STN	0.11 ±0.07	-2.98 ±0.49	-3.00 ±0.47	0.91 ±0.02
	vanilla attn.	0.14 ±0.01	-3.28 ±0.12	-3.26 ±0.18	0.88 ±0.01
	w/o continuity	0.12 ±0.03	-2.87 ±0.22	-2.95 ±0.22	0.91 ±0.01
	w/o diffusion	0.13 ±0.07	-2.94 ±0.60	-2.95 ±0.61	0.91 ±0.01
	w/o relaxation	0.07 ±0.18	-3.11 ±0.20	-3.27 ±0.09	0.91 ±0.01
	w/o advection	0.16 ±0.03	-3.24 ±0.44	-3.54 ±0.18	0.91 ±0.01
	w/o reaction	0.09 ±0.10	-2.49 ±0.29	-2.46 ±0.52	0.88 ±0.02
	w/o temporal	-3.02 ±0.29	-3.16 ±0.21	-3.02 ±0.25	0.85 ±0.01
Beijing	CyPhy-STN	-0.07 ±0.14	-1.40 ±0.43	-1.32 ±0.62	0.96 ±0.00
	vanilla attn.	-0.06 ±0.06	-1.47 ±0.42	-1.40 ±0.65	0.94 ±0.00
	w/o continuity	-0.10 ±0.09	-1.99 ±0.29	-1.70 ±0.14	0.95 ±0.00
	w/o diffusion	-0.06 ±0.11	-1.53 ±0.45	-1.50 ±0.47	0.95 ±0.00
	w/o relaxation	-0.15 ±0.14	-2.03 ±0.65	-1.82 ±0.43	1.01 ±0.02
	w/o advection	-0.11 ±0.06	-1.76 ±0.34	-1.66 ±0.38	0.95 ±0.00
	w/o reaction	-0.07 ±0.14	-1.41 ±0.42	-1.32 ±0.61	0.96 ±0.00
	w/o temporal	-1.31 ±0.44	-1.36 ±0.47	-1.31 ±0.73	0.95 ±0.00

shell contains only two reachable node pairs (see Appendix, Table 7). We therefore attribute this difference mainly to graph topology rather than to a distinct learnt propagation regime. Dirichlet ratios lie below one on all datasets ($\mathcal{D} = 0.80$–0.96). Each physics-message layer thus reduces variation between neighbouring representations. Together, these results characterise the physics stack as locally propagating and contractive.

The end-to-end Jacobians behave differently. On PEMS04 and PEMS08, sensitivity also drops sharply at the first hop ($\rho(2) = 0.10$ and 0.05). It then decays only slowly, retaining 52–57% per hop. On KnowAir and Beijing, it shows essentially no decay beyond the second hop ($\beta = +0.11$ and -0.07). On Beijing, even the first-hop drop is small ($\rho(2) = 0.87$). Neither physics-stack Jacobian shows such long-range sensitivity, and the node and context embeddings act node-wise. The long-range sensitivity must therefore enter through the temporal module. The *w/o temporal* ablation confirms this (Table 11). Without the temporal module, the end-to-end slope falls to that of the hidden-state Jacobian on all datasets. The locality of the physics stack itself remains essentially unchanged. The temporal module thus provides a long-range pathway that bypasses the physics-message stack. Local contraction within the stack therefore does not cause a loss of long-range sensitivity in the complete model.

Removing individual physical mechanisms shifts the hidden-state slope by up to 0.9. It ranges from -4.19 to -3.29 on PEMS04, from -4.76 to -4.18 on PEMS08, from -3.28 to -2.49 on

Table 12: Per-channel test errors on PEMS04 (masked MAE/SN MAE/RMSE/MASE, averaged over all 12 forecast steps). Values report mean ± standard deviation over available seeds (24, 33, 900, 2028, 2029). Best per row in bold; second best italic.

Channel	Metric	DLinear	Air-DualODE Multiv.	AirDDE	HGGFNO	STAE	DTR	CyPhy-van.	CyPhy	CyPhy w/o cont.	CyPhy w/o visc.	CyPhy w/o relax.	CyPhy w/o adv.	CyPhy w/o react.	CyPhy w/o temp.	Temporal
Flow (q)	MAE	28.2290 ± 0.08	21.6578 ± 0.06	23.5101 ± 1.25	18.8097 ± 0.10	18.9485 ± 0.09	19.2845 ± 0.30	18.8827 ± 0.02	18.7020 ± 0.01	18.7263 ± 0.06	*18.7009 ± 0.01*	**18.6912 ± 0.02**	18.7097 ± 0.02	18.8522 ± 0.03	19.4576 ± 0.05	19.5430 ± 0.05
	SN MAE	0.1807 ± 0.00	0.1386 ± 0.00	0.1505 ± 0.01	0.1204 ± 0.00	0.1213 ± 0.00	0.1234 ± 0.00	0.1209 ± 0.00	0.1197 ± 0.00	0.1199 ± 0.00	*0.1197 ± 0.00*	**0.1196 ± 0.00**	0.1197 ± 0.00	0.1207 ± 0.00	0.1245 ± 0.00	0.1251 ± 0.00
	RMSE	43.0063 ± 0.08	33.7843 ± 0.15	34.8066 ± 1.34	30.1788 ± 0.10	30.5823 ± 0.15	30.8307 ± 0.33	30.3639 ± 0.03	**30.1576 ± 0.01**	30.5536 ± 0.90	*30.1609 ± 0.02*	30.1799 ± 0.12	30.1705 ± 0.04	30.3968 ± 0.05	31.4547 ± 0.05	31.1334 ± 0.06
	MASE	0.6821 ± 0.00	0.5233 ± 0.00	0.5681 ± 0.03	0.4545 ± 0.00	0.4579 ± 0.00	0.4660 ± 0.01	0.4563 ± 0.00	0.4519 ± 0.00	0.4525 ± 0.00	*0.4519 ± 0.00*	**0.4517 ± 0.00**	0.4521 ± 0.00	0.4555 ± 0.00	0.4702 ± 0.00	0.4722 ± 0.00
Speed (v)	MAE	2.0300 ± 0.00	1.6553 ± 0.01	1.7098 ± 0.04	1.5911 ± 0.01	1.5668 ± 0.00	**1.5548 ± 0.01**	1.5746 ± 0.00	1.5635 ± 0.00	1.5620 ± 0.00	1.5647 ± 0.00	1.5642 ± 0.00	*1.5611 ± 0.00*	1.5716 ± 0.00	1.6013 ± 0.00	1.6276 ± 0.01
	SN MAE	0.2463 ± 0.00	0.2008 ± 0.00	0.2074 ± 0.01	0.1930 ± 0.00	0.1901 ± 0.00	**0.1886 ± 0.00**	0.1910 ± 0.00	0.1897 ± 0.00	0.1895 ± 0.00	0.1898 ± 0.00	0.1898 ± 0.00	*0.1894 ± 0.00*	0.1907 ± 0.00	0.1943 ± 0.00	0.1975 ± 0.00
	RMSE	4.5286 ± 0.00	3.6931 ± 0.04	3.5985 ± 0.05	3.6166 ± 0.01	3.6123 ± 0.03	**3.5269 ± 0.02**	3.5535 ± 0.01	3.5463 ± 0.01	*3.5404 ± 0.02*	3.5508 ± 0.01	3.5427 ± 0.01	3.5408 ± 0.01	3.5609 ± 0.01	3.5987 ± 0.02	3.6092 ± 0.02
	MASE	0.5802 ± 0.00	0.4731 ± 0.00	0.4887 ± 0.01	0.4548 ± 0.00	0.4478 ± 0.00	**0.4444 ± 0.00**	0.4501 ± 0.00	0.4469 ± 0.00	0.4465 ± 0.00	0.4472 ± 0.00	0.4471 ± 0.00	*0.4462 ± 0.00*	0.4492 ± 0.00	0.4577 ± 0.00	0.4652 ± 0.00
Occupancy (rho)	MAE	0.0103 ± 0.00	0.0081 ± 0.00	0.0085 ± 0.00	0.0074 ± 0.00	0.0075 ± 0.00	0.0075 ± 0.00	0.0074 ± 0.00	*0.0074 ± 0.00*	0.0074 ± 0.00	0.0074 ± 0.00	**0.0073 ± 0.00**	0.0074 ± 0.00	0.0074 ± 0.00	0.0076 ± 0.00	0.0078 ± 0.00
	SN MAE	0.2109 ± 0.00	0.1665 ± 0.00	0.1738 ± 0.01	0.1515 ± 0.00	0.1530 ± 0.00	0.1532 ± 0.00	0.1520 ± 0.00	*0.1509 ± 0.00*	0.1510 ± 0.00	0.1511 ± 0.00	**0.1509 ± 0.00**	0.1510 ± 0.00	0.1529 ± 0.00	0.1559 ± 0.00	0.1597 ± 0.00
	RMSE	0.0247 ± 0.00	0.0207 ± 0.00	0.0202 ± 0.00	0.0196 ± 0.00	0.0202 ± 0.00	0.0198 ± 0.00	0.0197 ± 0.00	0.0196 ± 0.00	0.0196 ± 0.00	0.0197 ± 0.00	*0.0196 ± 0.00*	**0.0196 ± 0.00**	0.0197 ± 0.00	0.0203 ± 0.00	0.0200 ± 0.00
	MASE	0.6424 ± 0.00	0.5073 ± 0.00	0.5293 ± 0.03	0.4615 ± 0.00	0.4660 ± 0.00	0.4666 ± 0.00	0.4631 ± 0.00	*0.4598 ± 0.00*	0.4599 ± 0.00	0.4603 ± 0.00	**0.4596 ± 0.00**	0.4599 ± 0.00	0.4657 ± 0.00	0.4750 ± 0.00	0.4865 ± 0.00

Table 13: Per-channel test errors on PEMS08 (masked MAE/SN MAE/RMSE/MASE, averaged over all 12 forecast steps). Values report mean ± standard deviation over available seeds (24, 33, 900, 2028, 2029). Best per row in bold; second best italic.

Channel	Metric	DLinear	Air-DualODE Multiv.	AirDDE	HGGFNO	STAE	DTR	CyPhy-van.	CyPhy	CyPhy w/o cont.	CyPhy w/o visc.	CyPhy w/o relax.	CyPhy w/o adv.	CyPhy w/o react.	CyPhy w/o temp.	Temporal
Flow (q)	MAE	22.7229 ± 0.02	16.3805 ± 0.11	18.2971 ± 0.40	14.6142 ± 0.13	**14.3693 ± 0.16**	14.5354 ± 0.05	14.4486 ± 0.06	*14.4230 ± 0.06*	14.4542 ± 0.02	14.4293 ± 0.05	14.4284 ± 0.08	14.4583 ± 0.11	14.5730 ± 0.11	14.9458 ± 0.09	15.4060 ± 0.04
	SN MAE	0.1562 ± 0.00	0.1126 ± 0.00	0.1258 ± 0.00	0.1004 ± 0.00	**0.0988 ± 0.00**	0.0999 ± 0.00	0.0993 ± 0.00	*0.0991 ± 0.00*	0.0993 ± 0.00	0.0992 ± 0.00	0.0992 ± 0.00	0.0994 ± 0.00	0.1002 ± 0.00	0.1027 ± 0.00	0.1059 ± 0.00
	RMSE	34.4808 ± 0.04	25.7790 ± 0.26	27.1214 ± 0.58	23.4531 ± 0.15	23.4444 ± 0.16	23.6355 ± 0.06	23.4725 ± 0.07	23.4774 ± 0.10	23.4910 ± 0.07	**23.4263 ± 0.04**	*23.4425 ± 0.12*	23.4961 ± 0.10	23.6646 ± 0.12	24.4284 ± 0.20	24.5467 ± 0.05
	MASE	0.6399 ± 0.00	0.4613 ± 0.00	0.5153 ± 0.01	0.4116 ± 0.00	**0.4047 ± 0.00**	0.4093 ± 0.00	0.4069 ± 0.00	*0.4062 ± 0.00*	0.4071 ± 0.00	0.4064 ± 0.00	0.4063 ± 0.00	0.4072 ± 0.00	0.4104 ± 0.00	0.4209 ± 0.00	0.4339 ± 0.00
Speed (v)	MAE	1.5962 ± 0.00	1.3301 ± 0.01	1.4305 ± 0.02	1.3221 ± 0.01	**1.2735 ± 0.00**	1.2750 ± 0.00	1.2841 ± 0.01	1.2834 ± 0.01	*1.2746 ± 0.01*	1.2784 ± 0.01	1.2784 ± 0.00	1.2799 ± 0.00	1.2900 ± 0.01	1.3019 ± 0.01	1.3154 ± 0.01
	SN MAE	0.2433 ± 0.00	0.2028 ± 0.00	0.2181 ± 0.00	0.2016 ± 0.00	**0.1942 ± 0.00**	0.1944 ± 0.00	0.1958 ± 0.00	0.1957 ± 0.00	*0.1943 ± 0.00*	0.1949 ± 0.00	0.1949 ± 0.00	0.1951 ± 0.00	0.1967 ± 0.00	0.1985 ± 0.00	0.2005 ± 0.00
	RMSE	3.6645 ± 0.00	3.1910 ± 0.06	3.2153 ± 0.06	3.2159 ± 0.04	3.2866 ± 0.01	3.2149 ± 0.01	3.2010 ± 0.03	3.2197 ± 0.03	*3.1767 ± 0.03*	3.1970 ± 0.02	3.2009 ± 0.01	3.2005 ± 0.03	3.2268 ± 0.03	3.2246 ± 0.02	**3.1599 ± 0.01**
	MASE	0.6033 ± 0.00	0.5027 ± 0.01	0.5407 ± 0.01	0.4997 ± 0.00	**0.4813 ± 0.00**	0.4819 ± 0.00	0.4854 ± 0.00	0.4851 ± 0.00	*0.4817 ± 0.00*	0.4832 ± 0.00	0.4832 ± 0.00	0.4837 ± 0.00	0.4876 ± 0.00	0.4921 ± 0.00	0.4972 ± 0.00
Occupancy (rho)	MAE	0.0089 ± 0.00	0.0068 ± 0.00	0.0075 ± 0.00	0.0065 ± 0.00	0.0064 ± 0.00	0.0064 ± 0.00	0.0064 ± 0.00	*0.0064 ± 0.00*	**0.0064 ± 0.00**	0.0064 ± 0.00	0.0064 ± 0.00	0.0064 ± 0.00	0.0065 ± 0.00	0.0066 ± 0.00	0.0067 ± 0.00
	SN MAE	0.1975 ± 0.00	0.1526 ± 0.00	0.1681 ± 0.00	0.1449 ± 0.00	0.1426 ± 0.00	0.1432 ± 0.00	0.1434 ± 0.00	*0.1425 ± 0.00*	**0.1425 ± 0.00**	0.1429 ± 0.00	0.1426 ± 0.00	0.1430 ± 0.00	0.1442 ± 0.00	0.1467 ± 0.00	0.1491 ± 0.00
	RMSE	0.0206 ± 0.00	0.0182 ± 0.00	0.0186 ± 0.00	*0.0178 ± 0.00*	0.0185 ± 0.00	0.0182 ± 0.00	0.0182 ± 0.00	0.0181 ± 0.00	0.0179 ± 0.00	0.0181 ± 0.00	0.0181 ± 0.00	0.0181 ± 0.00	0.0181 ± 0.00	0.0185 ± 0.00	**0.0176 ± 0.00**
	MASE	0.6134 ± 0.00	0.4740 ± 0.00	0.5222 ± 0.01	0.4499 ± 0.00	0.4428 ± 0.00	0.4449 ± 0.00	0.4455 ± 0.00	*0.4426 ± 0.00*	**0.4425 ± 0.00**	0.4438 ± 0.00	0.4430 ± 0.00	0.4442 ± 0.00	0.4480 ± 0.00	0.4558 ± 0.00	0.4630 ± 0.00

KnowAir, and from -2.03 to -1.41 on Beijing. The qualitative pattern is preserved in every case. The physics stack remains strongly local, and the end-to-end slope stays close to zero on KnowAir and Beijing (-0.15 to $+0.16$). Long-range end-to-end sensitivity is therefore not driven by a single physical mechanism.

Overall, CyPhy-STN combines strongly local, contractive propagation within the physics-message stack with long-range sensitivity provided by the temporal module.

A.2.3 Test Errors for every channel separately

The Tables 12, 13, 14 and 15 report the Mean Absolute Error (MAE), σ-normalised Mean Absolute Error (σn-MAE), Root Mean Square Error (RMSE), and the Mean Absolute Percentage Error (MASE) for every model, dataset and channel separately.

A.3 Benchmarking Model Adaptations

DLinear Zeng et al. (2023) The node and channel axes are flattened into $N * C$ independent series, each forecast separately by the shared linear maps, and the $[B, T, N * C]$ result is reshaped to $[B, T, N, C]$.

STAEformer Liu et al. (2023) The changes necessary on STAEformer were minimal. We generalize STAEformer to arbitrary multivariate inputs and outputs by decoupling physical channels from temporal covariates. The original input convention stores time-of-day and day-of-week at fixed channel indices, which collide with additional physical variables. We therefore place the C_{in} physical variables first and append the temporal covariates at indices C_{in} and $C_{\text{in}} + 1$. The physical channels $x[\ldots, : C_{\text{in}}]$ are jointly mapped to the input embedding by a linear projection, while the temporal covariates keep their original embedding layers and are concatenated afterwards. The resulting latent tensor matches the original architecture, so temporal and spatial attention operate unchanged along time and nodes, without per-variable branches or an additional channel-attention dimension. On the output side, each node's latent representation of shape $[T_{\text{in}}, D]$ is flattened and mapped by a single linear layer to $T_{\text{out}} * C_{\text{out}}$ values, reshaped to $[B, T_{\text{out}}, N, C_{\text{out}}]$. All output variables are thus predicted jointly from a shared spatio-temporal representation.

AirDualODE Tian et al. (2025) For Air-DualODE, the original architecture is designed for single-target pollutant forecasting, with meteorological variables used as auxiliary covariates. Extending it to multivariate forecasting therefore requires promoting several heterogeneous variables to jointly predicted state variables. In our adaptation, all C_{out} target channels are included in X, propagated through the diffusion–advection dynamics, and decoded jointly. This introduces a strong

Table 14: Per-channel test errors on KnowAir (masked MAE/SN MAE/RMSE/MASE, averaged over all 24 forecast steps). Values report mean ± standard deviation over available seeds (24, 33, 900, 2028, 2029). Best per row in bold; second best italic.

Channel	Metric	DLinear	Air-DualODE Multiv.	AirDDE	HGGFNO	STAE	DTR	CyPhy-van.	CyPhy	CyPhy w/o cont.	CyPhy w/o visc.	CyPhy w/o relax.	CyPhy w/o adv.	CyPhy w/o react.	CyPhy w/o temp.	Temporal
surface_pressure	MAE	332.8979 ± 4.01	364.3237 ± 6.59	394.6131 ± 5.66	314.7245 ± 3.98	382.7410 ± 30.52	381.7126 ± 26.71	307.2218 ± 5.23	*295.1130 ± 1.68*	295.8643 ± 2.43	295.5284 ± 3.92	301.2706 ± 2.98	295.9461 ± 3.38	296.3370 ± 2.15	327.1408 ± 3.71	**290.4814 ± 2.16**
	SN MAE	0.0580 ± 0.00	0.0635 ± 0.00	0.0549 ± 0.00	0.0547 ± 0.00	0.0667 ± 0.01	0.0665 ± 0.00	0.0536 ± 0.00	*0.0514 ± 0.00*	0.0516 ± 0.00	0.0515 ± 0.00	0.0525 ± 0.00	0.0516 ± 0.00	0.0517 ± 0.00	0.0570 ± 0.00	**0.0506 ± 0.00**
	RMSE	431.2605 ± 4.11	480.4710 ± 7.88	514.3583 ± 5.84	413.7413 ± 4.45	491.1291 ± 34.46	495.1446 ± 31.51	403.6472 ± 9.15	*387.4726 ± 3.55*	388.5794 ± 1.68	388.2381 ± 6.12	395.9604 ± 2.50	389.4019 ± 5.73	389.2376 ± 2.64	435.0639 ± 5.33	**382.7343 ± 2.63**
	MASE	1.1004 ± 0.01	1.2043 ± 0.02	1.3044 ± 0.02	1.0403 ± 0.01	1.2652 ± 0.10	1.2618 ± 0.09	1.0155 ± 0.02	*0.9755 ± 0.01*	0.9780 ± 0.01	0.9769 ± 0.01	0.9959 ± 0.01	0.9783 ± 0.01	0.9796 ± 0.01	1.0814 ± 0.01	**0.9602 ± 0.01**
boundary_layer_height	MAE	222.3982 ± 0.04	353.9902 ± 0.67	214.1980 ± 7.38	216.9244 ± 1.59	*199.7566 ± 2.32*	**198.1086 ± 1.32**	211.2987 ± 4.20	208.1733 ± 1.47	207.9025 ± 2.77	209.3270 ± 4.01	208.6763 ± 3.20	209.0035 ± 1.45	216.7552 ± 2.38	206.4615 ± 0.82	
	SN MAE	0.4313 ± 0.00	0.6865 ± 0.00	0.4154 ± 0.01	0.4207 ± 0.00	*0.3874 ± 0.00*	**0.3842 ± 0.00**	0.4098 ± 0.01	0.4037 ± 0.00	0.4032 ± 0.01	0.4059 ± 0.01	0.4047 ± 0.01	0.4029 ± 0.00	0.3976 ± 0.00	0.4204 ± 0.00	0.4004 ± 0.00
	RMSE	325.8265 ± 0.12	546.6904 ± 1.60	327.2743 ± 11.33	331.1130 ± 1.12	**303.3870 ± 1.97**	*303.9888 ± 1.47*	317.8295 ± 4.48	315.4184 ± 1.90	314.0401 ± 3.81	316.9449 ± 4.04	315.4141 ± 4.25	315.2291 ± 3.59	310.4370 ± 1.12	325.0766 ± 3.18	313.3371 ± 0.74
	MASE	0.9880 ± 0.00	1.5726 ± 0.00	0.9916 ± 0.03	0.9637 ± 0.01	*0.8874 ± 0.01*	**0.8801 ± 0.01**	0.9387 ± 0.02	0.9248 ± 0.01	0.9236 ± 0.01	0.9299 ± 0.02	0.9270 ± 0.01	0.9230 ± 0.01	0.9107 ± 0.01	0.9629 ± 0.01	0.9172 ± 0.00
pm25	MAE	19.6934 ± 0.00	19.2482 ± 0.16	19.1377 ± 0.22	18.9570 ± 0.08	18.7639 ± 0.17	18.7721 ± 0.28	**18.6954 ± 0.11**	18.7896 ± 0.13	18.7562 ± 0.07	18.7821 ± 0.13	18.7360 ± 0.07	18.7870 ± 0.11	*18.7282 ± 0.07*	19.2988 ± 0.02	18.8440 ± 0.03
	SN MAE	0.4079 ± 0.00	0.3987 ± 0.00	0.3964 ± 0.00	0.3926 ± 0.00	0.3886 ± 0.00	0.3888 ± 0.01	**0.3872 ± 0.00**	0.3892 ± 0.00	0.3885 ± 0.00	0.3890 ± 0.00	0.3880 ± 0.00	0.3891 ± 0.00	*0.3879 ± 0.00*	0.3997 ± 0.00	0.3903 ± 0.00
	RMSE	30.6762 ± 0.01	29.6158 ± 0.31	*28.7429 ± 0.09*	29.9831 ± 0.09	28.7433 ± 0.23	**28.6519 ± 0.22**	29.9277 ± 0.44	29.7552 ± 0.22	29.7392 ± 0.12	29.7926 ± 0.31	29.9245 ± 0.20	29.8137 ± 0.26	29.7439 ± 0.14	30.4323 ± 0.17	30.0772 ± 0.14
	MASE	0.7145 ± 0.00	0.6983 ± 0.01	0.6943 ± 0.01	0.6877 ± 0.00	0.6807 ± 0.01	0.6810 ± 0.01	**0.6782 ± 0.00**	0.6817 ± 0.00	0.6805 ± 0.00	0.6814 ± 0.00	0.6797 ± 0.00	0.6816 ± 0.00	*0.6794 ± 0.00*	0.7001 ± 0.00	0.6836 ± 0.00
relative_humidity_950	MAE	13.5452 ± 0.00	13.1337 ± 0.04	**11.8051 ± 0.17**	13.2548 ± 0.05	12.3889 ± 0.39	*12.0519 ± 0.14*	13.1852 ± 0.15	13.0115 ± 0.15	12.9731 ± 0.20	12.9173 ± 0.19	12.9909 ± 0.20	12.8980 ± 0.11	12.8817 ± 0.15	13.6334 ± 0.09	12.9776 ± 0.02
	SN MAE	0.5969 ± 0.00	0.5788 ± 0.00	**0.5202 ± 0.01**	0.5841 ± 0.00	0.5459 ± 0.02	*0.5311 ± 0.01*	0.5810 ± 0.01	0.5734 ± 0.01	0.5717 ± 0.01	0.5692 ± 0.01	0.5725 ± 0.01	0.5684 ± 0.01	0.5677 ± 0.01	0.6008 ± 0.00	0.5719 ± 0.00
	RMSE	17.3308 ± 0.01	16.6250 ± 0.03	**15.2592 ± 0.24**	16.8981 ± 0.06	15.9018 ± 0.46	*15.5524 ± 0.18*	17.0671 ± 0.18	16.8799 ± 0.21	16.8528 ± 0.20	16.7446 ± 0.25	16.8635 ± 0.23	16.7596 ± 0.12	16.7392 ± 0.21	17.5522 ± 0.10	16.8549 ± 0.04
	MASE	1.0019 ± 0.00	0.9715 ± 0.00	**0.8732 ± 0.01**	0.9804 ± 0.00	0.9164 ± 0.03	*0.8914 ± 0.01*	0.9753 ± 0.01	0.9624 ± 0.01	0.9596 ± 0.01	0.9554 ± 0.01	0.9609 ± 0.01	0.9540 ± 0.01	0.9528 ± 0.01	1.0084 ± 0.01	0.9599 ± 0.00
relative_humidity_975	MAE	12.7632 ± 0.00	13.0623 ± 0.04	**11.2314 ± 0.13**	12.5398 ± 0.04	11.7032 ± 0.37	*11.4050 ± 0.12*	12.4650 ± 0.17	12.2596 ± 0.11	12.2192 ± 0.21	12.1871 ± 0.20	12.2534 ± 0.21	12.1475 ± 0.12	12.1270 ± 0.14	12.8052 ± 0.09	12.2418 ± 0.03
	SN MAE	0.5745 ± 0.00	0.5880 ± 0.00	**0.5015 ± 0.01**	0.5634 ± 0.00	0.5268 ± 0.02	*0.5134 ± 0.01*	0.5611 ± 0.01	0.5518 ± 0.01	0.5500 ± 0.01	0.5486 ± 0.01	0.5516 ± 0.01	0.5468 ± 0.01	0.5459 ± 0.01	0.5764 ± 0.00	0.5510 ± 0.00
	RMSE	16.3735 ± 0.01	16.3825 ± 0.03	**14.5055 ± 0.20**	16.0148 ± 0.05	15.0350 ± 0.44	*14.7285 ± 0.16*	16.1353 ± 0.22	15.9112 ± 0.17	15.8813 ± 0.21	15.8034 ± 0.27	15.9141 ± 0.26	15.7952 ± 0.14	15.7608 ± 0.20	16.5128 ± 0.10	15.9251 ± 0.03
	MASE	0.9987 ± 0.00	1.0221 ± 0.00	**0.8788 ± 0.01**	0.9812 ± 0.00	0.9157 ± 0.03	*0.8924 ± 0.01*	0.9753 ± 0.01	0.9593 ± 0.01	0.9561 ± 0.02	0.9536 ± 0.02	0.9588 ± 0.02	0.9505 ± 0.01	0.9489 ± 0.01	1.0019 ± 0.01	0.9579 ± 0.00
k_index	MAE	10.6806 ± 0.00	9.7387 ± 0.06	**9.3205 ± 0.10**	10.3974 ± 0.05	9.8552 ± 0.35	*9.5977 ± 0.08*	10.2531 ± 0.12	10.1168 ± 0.09	10.1378 ± 0.08	10.0924 ± 0.12	10.1391 ± 0.16	10.0525 ± 0.06	10.0616 ± 0.10	10.5616 ± 0.07	10.2157 ± 0.05
	SN MAE	0.5144 ± 0.00	0.4691 ± 0.00	**0.4489 ± 0.00**	0.5008 ± 0.00	0.4747 ± 0.02	*0.4622 ± 0.00*	0.4938 ± 0.01	0.4873 ± 0.00	0.4883 ± 0.00	0.4861 ± 0.01	0.4883 ± 0.01	0.4842 ± 0.00	0.4846 ± 0.00	0.5087 ± 0.00	0.4920 ± 0.00
	RMSE	14.3405 ± 0.00	13.2968 ± 0.09	**12.7607 ± 0.09**	14.0304 ± 0.05	13.4150 ± 0.47	*13.0779 ± 0.15*	14.0374 ± 0.14	13.8853 ± 0.12	13.8712 ± 0.15	13.8539 ± 0.13	13.8992 ± 0.11	13.7555 ± 0.08	13.8792 ± 0.12	14.2580 ± 0.06	14.0209 ± 0.09
	MASE	0.8748 ± 0.00	0.7976 ± 0.01	**0.7634 ± 0.01**	0.8516 ± 0.00	0.8072 ± 0.03	*0.7860 ± 0.01*	0.8398 ± 0.01	0.8286 ± 0.01	0.8303 ± 0.01	0.8266 ± 0.01	0.8304 ± 0.01	0.8233 ± 0.00	0.8241 ± 0.01	0.8650 ± 0.01	0.8367 ± 0.00
v_component_of_wind_950	MAE	3.0531 ± 0.00	2.8763 ± 0.01	**2.7830 ± 0.02**	2.9917 ± 0.01	2.8791 ± 0.09	*2.8250 ± 0.03*	2.9183 ± 0.03	2.9654 ± 0.04	2.9733 ± 0.04	2.9535 ± 0.03	2.9573 ± 0.03	2.9437 ± 0.02	2.9458 ± 0.02	2.9923 ± 0.01	2.9761 ± 0.01
	SN MAE	0.6669 ± 0.00	0.6282 ± 0.00	**0.6079 ± 0.01**	0.6534 ± 0.00	0.6289 ± 0.02	*0.6170 ± 0.01*	0.6374 ± 0.01	0.6477 ± 0.01	0.6494 ± 0.01	0.6451 ± 0.01	0.6459 ± 0.01	0.6430 ± 0.00	0.6434 ± 0.00	0.6536 ± 0.00	0.6501 ± 0.00
	RMSE	4.3347 ± 0.00	3.9846 ± 0.01	**3.8859 ± 0.01**	4.2146 ± 0.02	3.9960 ± 0.10	*3.9545 ± 0.04*	4.1287 ± 0.04	4.2007 ± 0.06	4.2061 ± 0.05	4.1824 ± 0.05	4.1989 ± 0.05	4.1720 ± 0.04	4.1771 ± 0.03	4.2573 ± 0.03	4.2041 ± 0.01
	MASE	0.8195 ± 0.00	0.7721 ± 0.00	**0.7470 ± 0.01**	0.8030 ± 0.00	0.7728 ± 0.02	*0.7583 ± 0.01*	0.7833 ± 0.01	0.7960 ± 0.01	0.7981 ± 0.01	0.7928 ± 0.01	0.7938 ± 0.01	0.7902 ± 0.01	0.7907 ± 0.01	0.8032 ± 0.00	0.7989 ± 0.00
2m_dewpoint_temperature	MAE	3.0946 ± 0.00	2.5592 ± 0.02	**2.4764 ± 0.02**	2.8366 ± 0.02	2.6023 ± 0.10	*2.5494 ± 0.03*	2.6885 ± 0.02	2.6853 ± 0.05	2.6715 ± 0.03	2.6893 ± 0.03	2.7228 ± 0.04	2.6540 ± 0.01	2.6877 ± 0.03	2.9592 ± 0.04	2.7334 ± 0.01
	SN MAE	0.2676 ± 0.00	0.2213 ± 0.00	**0.2141 ± 0.00**	0.2453 ± 0.00	0.2250 ± 0.01	*0.2204 ± 0.00*	0.2325 ± 0.00	0.2322 ± 0.00	0.2310 ± 0.00	0.2325 ± 0.00	0.2354 ± 0.00	0.2295 ± 0.00	0.2324 ± 0.00	0.2559 ± 0.00	0.2364 ± 0.00
	RMSE	4.1750 ± 0.00	3.5512 ± 0.02	**3.4519 ± 0.04**	3.9280 ± 0.03	3.5918 ± 0.12	*3.5227 ± 0.05*	3.7569 ± 0.02	3.7634 ± 0.08	3.7365 ± 0.05	3.7537 ± 0.03	3.8039 ± 0.04	3.7169 ± 0.02	3.7804 ± 0.05	4.0531 ± 0.05	3.8213 ± 0.01
	MASE	1.1438 ± 0.00	0.9459 ± 0.01	**0.9153 ± 0.01**	1.0484 ± 0.01	0.9619 ± 0.04	*0.9423 ± 0.01*	0.9937 ± 0.01	0.9925 ± 0.02	0.9874 ± 0.01	0.9940 ± 0.01	1.0064 ± 0.01	0.9934 ± 0.01	1.0017 ± 0.02	1.0983 ± 0.00	1.0102 ± 0.00
u_component_of_wind_950	MAE	2.5862 ± 0.00	2.5383 ± 0.01	**2.4767 ± 0.04**	2.5330 ± 0.01	2.5659 ± 0.06	*2.5152 ± 0.03*	2.5876 ± 0.04	2.6006 ± 0.04	2.6229 ± 0.04	2.5886 ± 0.04	2.6020 ± 0.03	2.5874 ± 0.02	2.5919 ± 0.02	2.6210 ± 0.01	2.5788 ± 0.02
	SN MAE	0.6285 ± 0.00	0.6168 ± 0.00	**0.6019 ± 0.01**	0.6155 ± 0.00	0.6235 ± 0.01	*0.6112 ± 0.01*	0.6288 ± 0.01	0.6320 ± 0.01	0.6374 ± 0.01	0.6291 ± 0.01	0.6323 ± 0.01	0.6288 ± 0.01	0.6299 ± 0.01	0.6369 ± 0.00	0.6267 ± 0.00
	RMSE	3.6528 ± 0.00	3.4903 ± 0.02	**3.4025 ± 0.06**	3.5641 ± 0.01	3.5329 ± 0.08	*3.4831 ± 0.05*	3.6288 ± 0.06	3.6419 ± 0.05	3.6755 ± 0.05	3.6253 ± 0.05	3.6438 ± 0.03	3.6258 ± 0.03	3.6315 ± 0.03	3.6770 ± 0.00	3.6338 ± 0.02
	MASE	0.8341 ± 0.00	0.8187 ± 0.00	**0.7988 ± 0.01**	0.8170 ± 0.00	0.8276 ± 0.02	*0.8112 ± 0.01*	0.8346 ± 0.01	0.8388 ± 0.01	0.8460 ± 0.01	0.8349 ± 0.01	0.8392 ± 0.01	0.8345 ± 0.01	0.8360 ± 0.01	0.8454 ± 0.00	0.8318 ± 0.00
temperature_950	MAE	2.8953 ± 0.00	2.5351 ± 0.04	**2.2904 ± 0.04**	2.6028 ± 0.04	2.4148 ± 0.11	*2.3700 ± 0.06*	2.4825 ± 0.01	2.5097 ± 0.05	2.4835 ± 0.05	2.4928 ± 0.01	2.5122 ± 0.05	2.4726 ± 0.01	2.4920 ± 0.02	2.7235 ± 0.04	2.5052 ± 0.00
	SN MAE	0.2903 ± 0.00	0.2542 ± 0.00	**0.2296 ± 0.00**	0.2609 ± 0.00	0.2421 ± 0.01	*0.2376 ± 0.01*	0.2489 ± 0.00	0.2516 ± 0.01	0.2490 ± 0.00	0.2499 ± 0.00	0.2519 ± 0.01	0.2479 ± 0.00	0.2498 ± 0.00	0.2730 ± 0.00	0.2512 ± 0.00
	RMSE	3.8214 ± 0.00	3.3076 ± 0.04	**2.9309 ± 0.05**	3.4765 ± 0.05	3.1718 ± 0.13	*3.1303 ± 0.06*	3.2931 ± 0.02	3.3399 ± 0.09	3.3103 ± 0.07	3.3110 ± 0.02	3.3338 ± 0.06	3.2931 ± 0.03	3.3297 ± 0.03	3.6089 ± 0.04	3.3482 ± 0.01
	MASE	1.1554 ± 0.00	1.0116 ± 0.02	**0.9140 ± 0.02**	1.0386 ± 0.02	0.9636 ± 0.05	*0.9457 ± 0.02*	0.9906 ± 0.01	1.0015 ± 0.02	0.9910 ± 0.02	0.9947 ± 0.01	1.0025 ± 0.02	0.9867 ± 0.01	0.9944 ± 0.01	1.0868 ± 0.02	0.9997 ± 0.00
temperature_925	MAE	2.8764 ± 0.00	2.4511 ± 0.04	**2.2647 ± 0.05**	2.5773 ± 0.04	2.4000 ± 0.12	*2.3516 ± 0.06*	2.4509 ± 0.01	2.4779 ± 0.06	2.4593 ± 0.05	2.4630 ± 0.01	2.4905 ± 0.05	2.4413 ± 0.02	2.4604 ± 0.03	2.6910 ± 0.04	2.4550 ± 0.02
	SN MAE	0.2915 ± 0.00	0.2484 ± 0.00	**0.2296 ± 0.00**	0.2612 ± 0.00	0.2433 ± 0.01	*0.2384 ± 0.01*	0.2484 ± 0.00	0.2512 ± 0.01	0.2493 ± 0.00	0.2497 ± 0.00	0.2524 ± 0.01	0.2474 ± 0.00	0.2494 ± 0.00	0.2728 ± 0.00	0.2488 ± 0.00
	RMSE	3.8035 ± 0.00	3.2276 ± 0.05	**3.0036 ± 0.07**	3.4507 ± 0.05	3.1531 ± 0.13	*3.1108 ± 0.06*	3.2562 ± 0.02	3.3014 ± 0.09	3.2772 ± 0.07	3.2730 ± 0.02	3.3084 ± 0.06	3.2521 ± 0.04	3.2908 ± 0.04	3.5717 ± 0.04	3.2915 ± 0.02
	MASE	1.1529 ± 0.00	0.9825 ± 0.02	**0.9077 ± 0.02**	1.0330 ± 0.02	0.9620 ± 0.05	*0.9425 ± 0.02*	0.9824 ± 0.01	0.9932 ± 0.02	0.9857 ± 0.02	0.9872 ± 0.00	0.9982 ± 0.02	0.9785 ± 0.01	0.9862 ± 0.01	1.0786 ± 0.02	0.9840 ± 0.01
100m_v_component_of_wind	MAE	2.5750 ± 0.00	2.4317 ± 0.01	**2.3676 ± 0.03**	2.5215 ± 0.01	2.4294 ± 0.07	*2.3930 ± 0.03*	2.4711 ± 0.02	2.5042 ± 0.03	2.5111 ± 0.02	2.4932 ± 0.02	2.4985 ± 0.02	2.4850 ± 0.01	2.4867 ± 0.01	2.5300 ± 0.01	2.5239 ± 0.01
	SN MAE	0.7084 ± 0.00	0.6690 ± 0.00	**0.6514 ± 0.01**	0.6937 ± 0.00	0.6684 ± 0.02	*0.6584 ± 0.01*	0.6799 ± 0.01	0.6890 ± 0.01	0.6909 ± 0.01	0.6859 ± 0.01	0.6874 ± 0.01	0.6837 ± 0.00	0.6842 ± 0.00	0.6961 ± 0.00	0.6944 ± 0.00
	RMSE	3.4231 ± 0.00	3.1918 ± 0.01	**3.1371 ± 0.04**	3.3361 ± 0.01	3.2007 ± 0.08	*3.1753 ± 0.04*	3.2927 ± 0.03	3.3465 ± 0.05	3.3526 ± 0.02	3.3287 ± 0.04	3.3359 ± 0.03	3.3232 ± 0.02	3.3283 ± 0.02	3.3713 ± 0.02	3.3567 ± 0.01
	MASE	0.8418 ± 0.00	0.7949 ± 0.00	**0.7740 ± 0.01**	0.8242 ± 0.00	0.7942 ± 0.02	*0.7822 ± 0.01*	0.8078 ± 0.01	0.8186 ± 0.01	0.8209 ± 0.01	0.8150 ± 0.01	0.8168 ± 0.01	0.8123 ± 0.00	0.8129 ± 0.00	0.8271 ± 0.00	0.8273 ± 0.00
2m_temperature	MAE	2.5910 ± 0.00	3.1122 ± 0.02	*2.1566 ± 0.05*	2.4226 ± 0.02	2.1760 ± 0.09	**2.1406 ± 0.02**	2.2592 ± 0.01	2.2719 ± 0.03	2.2592 ± 0.02	2.2713 ± 0.02	2.2823 ± 0.04	2.2490 ± 0.01	2.2449 ± 0.02	2.4763 ± 0.03	2.2827 ± 0.02
	SN MAE	0.2495 ± 0.00	0.2997 ± 0.00	*0.2077 ± 0.00*	0.2333 ± 0.00	0.2095 ± 0.01	**0.2061 ± 0.00**	0.2175 ± 0.00	0.2188 ± 0.00	0.2175 ± 0.00	0.2187 ± 0.00	0.2198 ± 0.00	0.2165 ± 0.00	0.2161 ± 0.00	0.2384 ± 0.00	0.2198 ± 0.00
	RMSE	3.3973 ± 0.00	4.0016 ± 0.03	*2.8509 ± 0.08*	3.2267 ± 0.03	2.8615 ± 0.10	**2.8239 ± 0.02**	2.9898 ± 0.02	3.0042 ± 0.04	2.9903 ± 0.03	3.0007 ± 0.02	3.0167 ± 0.04	2.9763 ± 0.03	2.9839 ± 0.03	3.2673 ± 0.03	3.0422 ± 0.02
	MASE	1.1876 ± 0.00	1.4264 ± 0.01	*0.9885 ± 0.02*	1.1104 ± 0.01	0.9973 ± 0.04	**0.9811 ± 0.01**	1.0355 ± 0.01	1.0413 ± 0.01	1.0355 ± 0.01	1.0410 ± 0.01	1.0460 ± 0.02	1.0308 ± 0.01	1.0289 ± 0.01	1.1350 ± 0.01	1.0463 ± 0.01
100m_u_component_of_wind	MAE	2.0625 ± 0.00	2.0642 ± 0.01	2.0257 ± 0.03	*2.0256 ± 0.01*	2.0551 ± 0.03	**2.0220 ± 0.02**	2.0764 ± 0.02	2.0796 ± 0.02	2.0958 ± 0.02	2.0730 ± 0.02	2.0770 ± 0.02	2.0737 ± 0.01	2.0732 ± 0.01	2.0991 ± 0.01	2.0551 ± 0.01
	SN MAE	0.6720 ± 0.00	0.6726 ± 0.00	0.6601 ± 0.01	*0.6600 ± 0.00*	0.6696 ± 0.01	**0.6588 ± 0.01**	0.6766 ± 0.01	0.6776 ± 0.01	0.6829 ± 0.01	0.6755 ± 0.01	0.6768 ± 0.01	0.6757 ± 0.00	0.6755 ± 0.00	0.6840 ± 0.00	0.6696 ± 0.00
	RMSE	2.7285 ± 0.00	2.6870 ± 0.00	**2.6392 ± 0.04**	2.6750 ± 0.01	2.6788 ± 0.04	*2.6499 ± 0.03*	2.7390 ± 0.04	2.7417 ± 0.03	2.7600 ± 0.03	2.7345 ± 0.03	2.7380 ± 0.02	2.7363 ± 0.02	2.7351 ± 0.02	2.7685 ± 0.00	2.7189 ± 0.01
	MASE	0.8258 ± 0.00	0.8265 ± 0.00	0.8111 ± 0.01	*0.8110 ± 0.00*	0.8229 ± 0.01	**0.8096 ± 0.01**	0.8314 ± 0.01	0.8327 ± 0.01	0.8391 ± 0.01	0.8300 ± 0.01	0.8316 ± 0.01	0.8303 ± 0.01	0.8301 ± 0.01	0.8405 ± 0.00	0.8229 ± 0.00
vertical_velocity_950	MAE	0.1146 ± 0.00	0.1206 ± 0.00	0.1164 ± 0.00	0.1129 ± 0.00	0.1104 ± 0.00	0.1110 ± 0.00	0.1117 ± 0.00	0.1105 ± 0.00	0.1106 ± 0.00	0.1106 ± 0.00	0.1105 ± 0.00	*0.1104 ± 0.00*	**0.1100 ± 0.00**	0.1141 ± 0.00	0.1105 ± 0.00
	SN MAE	0.6499 ± 0.00	0.6837 ± 0.00	0.6600 ± 0.00	0.6399 ± 0.00	0.6260 ± 0.00	0.6292 ± 0.00	0.6334 ± 0.00	0.6266 ± 0.00	0.6268 ± 0.00	0.6270 ± 0.00	0.6263 ± 0.00	*0.6258 ± 0.00*	**0.6239 ± 0.00**	0.6470 ± 0.00	0.6265 ± 0.00
	RMSE	0.1667 ± 0.00	0.1725 ± 0.00	0.1674 ± 0.00	0.1638 ± 0.00	0.1615 ± 0.00	0.1624 ± 0.00	0.1626 ± 0.00	0.1613 ± 0.00	0.1614 ± 0.00	0.1613 ± 0.00	0.1611 ± 0.00	*0.1610 ± 0.00*	**0.1608 ± 0.00**	0.1650 ± 0.00	0.1619 ± 0.00
	MASE	0.8183 ± 0.00	0.8609 ± 0.00	0.8311 ± 0.00	0.8061 ± 0.00	0.7882 ± 0.00	0.7923 ± 0.00	0.7976 ± 0.00	0.7890 ± 0.00	0.7892 ± 0.00	0.7895 ± 0.00	0.7885 ± 0.00	*0.7880 ± 0.00*	**0.7856 ± 0.00**	0.8067 ± 0.00	0.8147 ± 0.00
specific_humidity_950	MAE	0.0015 ± 0.00	0.0013 ± 0.00	**0.0013 ± 0.00**	0.0014 ± 0.00	0.0014 ± 0.00	*0.0013 ± 0.00*	0.0014 ± 0.00	0.0014 ± 0.00	0.0014 ± 0.00	0.0014 ± 0.00	0.0014 ± 0.00	0.0014 ± 0.00	0.0014 ± 0.00	0.0015 ± 0.00	0.0014 ± 0.00
	SN MAE	0.3036 ± 0.00	0.2604 ± 0.00	**0.2474 ± 0.00**	0.2835 ± 0.00	0.2655 ± 0.01	*0.2558 ± 0.00*	0.2681 ± 0.00	0.2688 ± 0.00	0.2679 ± 0.00	0.2694 ± 0.00	0.2706 ± 0.00	0.2664 ± 0.00	0.2682 ± 0.00	0.2883 ± 0.00	0.2708 ± 0.00
	RMSE	0.0021 ± 0.00	0.0018 ± 0.00	**0.0017 ± 0.00**	0.0019 ± 0.00	0.0018 ± 0.00	*0.0018 ± 0.00*	0.0019 ± 0.00	0.0019 ± 0.00	0.0019 ± 0.00	0.0019 ± 0.00	0.0019 ± 0.00	0.0019 ± 0.00	0.0019 ± 0.00	0.0020 ± 0.00	0.0019 ± 0.00
	MASE	1.1138 ± 0.00	0.9554 ± 0.01	**0.9078 ± 0.01**	1.0401 ± 0.01	0.9739 ± 0.05	*0.9383 ± 0.01*	0.9834 ± 0.01	0.9861 ± 0.01	0.9827 ± 0.01	0.9882 ± 0.01	0.9928 ± 0.01	0.9773 ± 0.00	0.9839 ± 0.01	1.0577 ± 0.01	0.9935 ± 0.00
total_precipitation	MAE	0.0006 ± 0.00	0.0007 ± 0.00	0.0004 ± 0.00	**0.0004 ± 0.00**	0.0004 ± 0.00	0.0007 ± 0.00	0.0004 ± 0.00	0.0004 ± 0.00	0.0004 ± 0.00	0.0004 ± 0.00	0.0004 ± 0.00	0.0004 ± 0.00	0.0004 ± 0.00	*0.0004 ± 0.00*	0.0004 ± 0.00
	SN MAE	1.2560 ± 0.00	1.4278 ± 0.00	0.7556 ± 0.59	**0.7327 ± 0.51**	0.7540 ± 0.59	1.3954 ± 0.00	0.7478 ± 0.56	0.7497 ± 0.56	0.7543 ± 0.57	0.7504 ± 0.56	0.7503 ± 0.56	0.7549 ± 0.57	0.7506 ± 0.56	*0.7382 ± 0.52*	0.7435 ± 0.56
	RMSE	0.0012 ± 0.00	0.0013 ± 0.00	0.0008 ± 0.00	**0.0008 ± 0.00**	0.0009 ± 0.00	0.0013 ± 0.00	0.0009 ± 0.00	0.0009 ± 0.00	0.0009 ± 0.00	0.0009 ± 0.00	0.0009 ± 0.00	0.0009 ± 0.00	0.0009 ± 0.00	*0.0008 ± 0.00*	0.0009 ± 0.00
	MASE	3.3117 ± 0.00	3.7647 ± 0.00	1.9922 ± 1.55	**1.9320 ± 1.34**	1.9882 ± 1.56	3.6792 ± 0.01	1.9717 ± 1.48	1.9766 ± 1.48	1.9888 ± 1.50	1.9786 ± 1.48	1.9784 ± 1.48	1.9904 ± 1.50	1.9792 ± 1.48	*1.9464 ± 1.37*	1.9604 ± 1.47
vorticity_950	MAE	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	*0.0001 ± 0.00*	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	**0.0001 ± 0.00**	0.0001 ± 0.00	0.0001 ± 0.00
	SN MAE	2.4737 ± 0.00	2.8372 ± 0.00	2.4877 ± 1.15	2.4877 ± 1.15	2.4877 ± 1.15	2.6664 ± 0.00	2.4877 ± 1.15	2.4877 ± 1.15	2.4877 ± 1.15	2.4877 ± 1.15	2.4877 ± 1.15	2.4877 ± 1.15	**1.3866 ± 1.12**	1.4018 ± 1.03	1.4105 ± 1.05
	RMSE	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	*0.0001 ± 0.00*	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	0.0001 ± 0.00	**0.0001 ± 0.00**	0.0001 ± 0.00	0.0001 ± 0.00
	MASE	2.9809 ± 0.00	3.4190 ± 0.00	1.7927 ± 1.39	1.7372 ± 1.30	1.7502 ±

Table 16: Cross-domain σ-NMAE average rank (Table 4) before and after excluding the low-scale channels of Table 17. Traffic ranks are unchanged. Lower is better; negative Δ means the model moves up.

Model	Original	Revised	Δ	KnowAir rank	Beijing rank
CyPhy-STN	2.00	2.25	+0.25	4.00	2.00
DTRformer	3.75	2.50	-1.25	2.00	1.00
STAEformer	2.25	2.75	+0.50	3.00	5.00
HGGFNO	3.00	3.75	+0.75	5.00	3.00
AirDDE	5.00	5.00	0.00	1.00	7.00
Air-DualODE Multiv.	5.50	5.25	-0.25	7.00	4.00
DLinear	6.50	6.50	0.00	6.00	6.00

reshaped by the adapter to the common $[B, T_{\text{out}}, N, C_{\text{out}}]$ format for consistent per-channel evaluation.

AirDDE Wu et al. (2026) For AirDDE, the original decoder operates sequentially over the forecast horizon: at each future step, it produces a latent state of shape $[B, N, D_{\text{dec}}]$, applies the same output projection, and stacks the per-step predictions. The multivariate adaptation therefore only requires widening this projection from a scalar output to C_{out} channels, yielding $[B, N, C_{\text{out}}]$ per step and a final tensor of shape $[B, T_{\text{out}}, N, C_{\text{out}}]$. However, the underlying physical branch remains explicitly univariate: only a single `state_index` is evolved by the physical ODE, while the remaining variables are treated as covariates. Future covariates are not predicted explicitly but approximated from the observed history by repeating the last value. Thus, most output variables are produced only through the decoder projection instead of being modelled as dynamical states themselves, making the architecture poorly matched to joint multivariate forecasting.

DTRformer Chen et al. (2025) The adaptation in DTRformer is done equivalently to the one of STAEformer. The output tensor has shape $[B, T_{\text{out}}, N, C_{\text{out}}]$. With the mixed projection, the latent representation is first reshaped from $[B, T_{\text{in}}, N, D]$ to $[B, N, T_{\text{in}} * D]$. A linear layer then projects each node to $T_{\text{out}} * C_{\text{out}}$ values, giving the intermediate shape $[B, N, T_{\text{out}} * C_{\text{out}}]$, which is reshaped and transposed to the final output layout.

HG-GFNO Hosseini et al. (2026) The single-channel reference block in the original HG-GFNO implementation is applied independently to each input channel using shared weights, then the forecasts are stacked into the standard $[B, T, N, C]$ output. This preserves the reference architecture and spatial modelling but introduces no cross-channel interaction. Parameter count is independent of the number of channels, while computation scales linearly with it.

A.4 Effect of Low-Scale and Intermittent Channels on Normalised Aggregates

Normalised aggregates weight all channels equally, so channels with very small scale or intermittent behaviour can dominate them: normalisation inflates their errors relative to their absolute magnitude. To assess this sensitivity, Table 17 reports the aggregate σ-NMAE and MASE with and without the affected channels (total precipitation and vorticity on KnowAir, rain on Beijing). Excluding them reduces the aggregates by 10–67%. The largest reduction occurs for MASE on Beijing (64–67%), where rain is intermittent. The effect is similar across models on Beijing but uneven on KnowAir, where the reduction ranges from 14% (CyPhy-STN, HGGFNO) to 30% (DTRformer). The cross-domain ranking changes accordingly (Table 16): DTRformer moves from fifth to second place (average rank 3.75 to 2.50), while CyPhy-STN remains first (2.00 to 2.25). The small drop for CyPhy-STN comes entirely from KnowAir, where DTRformer moves ahead of it and it falls from third to fourth. Its traffic and Beijing ranks are unchanged. The excluded channels are not forecasting targets, so target-channel results are unaffected.

Table 17: Effect of low-scale non-target channels on equal-channel-weight normalised aggregates. Scales are inferred from MAE/σ-NMAE and MAE/MASE ratios. Channel errors give the range across seeds. Aggregate columns show all channels $\rightarrow$ without the listed channels, with the relative reduction in brackets.

Dataset	Channels	Model	σ scale	MASE scale	Channel σ-NMAE	Channel MASE	Aggregate σ-NMAE	Aggregate MASE
KnowAir	total_precipitation, vorticity_950	DLinear	$4.8 \cdot 10^{-5}$–$5.2 \cdot 10^{-4}$	$4.0 \cdot 10^{-5}$–$2.0 \cdot 10^{-4}$	1.26–2.47	2.98–3.31	0.61→0.46 (26%)	1.21→0.97 (20%)
		Air-DualODE Multiv.			1.43–2.84	3.41–3.77	0.64→0.46 (29%)	1.27→0.98 (23%)
		HGGFNO			0.36–2.63	0.78–3.41	0.51→0.44 (14%)	1.03→0.93 (10%)
		STAEformer			0.32–2.68	0.77–3.70	0.50→0.42 (15%)	1.00→0.90 (11%)
		DTRformer			1.39–2.68	3.20–3.69	0.60→0.42 (30%)	1.17→0.88 (24%)
		AirDDE			0.32–2.78	0.78–3.71	0.49→0.42 (16%)	0.99→0.88 (11%)
		CyPhy-STN			0.34–2.53	0.77–3.60	0.51→0.43 (14%)	1.00→0.90 (10%)
Beijing	rain	DLinear	0.72	0.10	2.74	18.96–18.97	0.65→0.47 (27%)	2.14→0.74 (66%)
		Air-DualODE Multiv.			2.76	19.09–19.10	0.64→0.46 (28%)	2.12→0.71 (67%)
		HGGFNO			2.64–2.67	18.30–18.47	0.62→0.45 (27%)	2.05→0.69 (66%)
		STAEformer			2.74–2.76	18.99–19.08	0.64→0.46 (28%)	2.13→0.72 (66%)
		DTRformer			2.75–2.76	19.05–19.13	0.62→0.45 (28%)	2.11→0.69 (67%)
		AirDDE			2.68–2.70	18.56–18.66	0.66→0.49 (25%)	2.16→0.79 (64%)
		CyPhy-STN			2.69–2.71	18.59–18.74	0.62→0.45 (28%)	2.07→0.69 (67%)