PAS^2: Physics-Anchored Spectral Reasoning for Air Quality Forecasting

Haofeng Ying[1,2] **Wenbin Lu**[1,3] **Wenxin Shen**[1] **Junnan Xu**[1] **Jianwei Zheng**[1]*
[1]Zhejiang University of Technology
[2]Zhejiang University
[3]Peking University

Abstract

Given the profound impacts on human health and ecosystems, air pollution necessitates robust quality prediction to inform policy-making. Traditional physics-based forecasting relies on idealized closed-system assumptions, whereas data-driven approaches often lack interpretability. Recent hybrid methods, however, neglect data homogeneity, resulting in redundancy, parameter inefficiency, and gradient conflicts. Moreover, high-concentration regions inherently exhibit drastic fluctuations, making it difficult to explicitly distinguish them from the overall pollution field, especially when using spatial information alone. To address these limitations, a Physics-Anchored Spatial-Spectral (PAS^2) paradigm is proposed. Specifically, a collaboration scheme is firstly elaborated, within which an independently supervised physical estimator captures deterministic trends as a predictive anchor, while its inferred physical regime conditions the neural branch to learn complementary fluctuations. Building on this, we equip the neural branch with a Spatial–Spectral (S^2) Reasoner that integrates graph operators to model local spatial diffusion and spectral operators to explicitly capture the drastic fluctuations in high-concentration regions. Experiments on real-world datasets show that PAS^2 consistently improves over strong data-driven and physics-guided baselines, especially under volatile sudden-change scenarios. Codes will be released at GitHub.

1 Introduction

Air quality forecasting is critical for public health policy and environmental regulation, yet it remains a daunting computational challenge due to the interplay of nonlinear transport dynamics, multi-scale meteorological forcing, and heterogeneous emission patterns (Azimi & Rahman, 2024; Cheng et al., 2024). Pollutant evolution combines structured diffusion–advection transport with abrupt, spatially non-local, broadband fluctuations. To tackle these complexities, recent research has increasingly pivoted towards Physics-Guided Learning (PGL), aiming to reconcile the interpretability of physical laws with the data-driven flexibility of deep learning by integrating differential equations into neural architectures (Seo et al., 2019; Guen & Thome, 2020; Hettige et al., 2024; Tian et al., 2025; Shi et al., 2025). However, despite advances, existing methods face two fundamental bottlenecks regarding architectural efficiency and signal representation. In this work, we propose the PAS^2 Strategy to address these limitations through a strict separation of concerns.

First, existing hybrid architectures often suffer from structural redundancy and cross-branch optimization interference. In practice, many PGL models adopt a “parallel dual-track" design in which the physical and neural branches operate as peer predictors. Without explicit role separation, the high-capacity neural branch may relearn dominant trends in the underlying pollutant evolution already

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

captured by the physical component, leading to parameter inefficiency and weakening the effect of the physical prior (Wang et al., 2025; Hwang & Lim, 2024; Yang et al., 2025). To reduce this architectural redundancy, we establish a Physics-Anchored Collaboration scheme. Unlike entangled parallel designs, our strategy employs a physical estimator to capture deterministic macroscopic trends, which serve as a detached anchor for the neural branch. This encourages the neural component to focus on complementary residual fluctuations rather than rediscovering basic transport patterns, thereby reducing direct cross-branch interference. Moreover, the physical trajectory and learned transport parameters condition the regime-adaptive expert routing, allowing the physical estimator to guide residual reasoning beyond providing a predictive baseline.

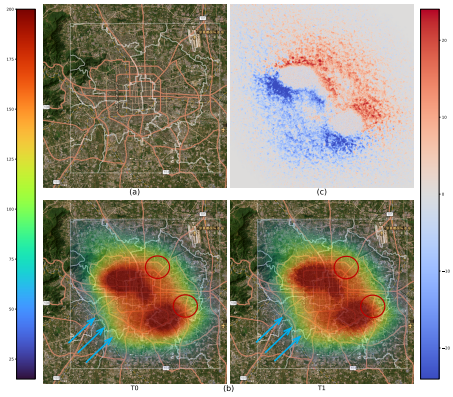


Figure 1: Visualization of $PM_{2.5}$ dynamics. (a) Satellite imagery. (b) $PM_{2.5}$ distributions at two timestamps. (c) Differential concentration map.

In addition, conventional spatial representations fail to capture the high-frequency volatility inherent to high-pollution regions. As illustrated in Fig. 1, while spatial pollutant distributions may appear continuous, their differential dynamics reveal a stark contrast: high-concentration regions exhibit drastic fluctuations (accumulation and depletion) that are significantly amplified compared to low-concentration areas. Existing solutions, which are based solely on spatial aggregators, typically prioritize local smoothness in the learned representations of neighboring monitoring stations. This inevitably "over-smooths" the distinct high-frequency events and obscures critical peak dynamics (Li et al., 2018a; Oono & Suzuki, 2020). To bridge this gap, we equip the neural branch with the S^2 Reasoner. This module synergizes graph operators for local diffusion with Spectral Operators equipped with FFT transformer. By simultaneously leveraging the spatial (Li et al., 2018b; Yu et al., 2018; Jiang et al., 2023b) and frequency domains (Li et al., 2021; Cao et al., 2020; Yi et al., 2023; Jiang et al., 2023a), this mechanism explicitly recovers the rapid regime shifts characteristic of high-concentration regions, allowing the model to distinguish volatile signals that are otherwise lost in purely spatial treatments. The main contributions of this work are threefold.

- We propose a physics-anchored residual framework for air quality forecasting. By using the physical estimator as a detached anchor and a conditioning signal, the neural branch is encouraged to model complementary residual fluctuations, reducing functional redundancy between physical and neural predictors.
- We introduce S^2 Reasoner to bridge the gap between spatial-frequency domains. By coupling graph-based spatial modeling with spectral operators, this mechanism explicitly resolves the distinct transport dynamics exhibited across varying concentration regimes.
- Extensive experiments on real-world datasets demonstrate that our framework surpasses existing methods in achieving state-of-the-art results, exhibiting superior robustness and generalization capabilities.

2 Related Work

2.1 Paradigms of Air Quality Prediction

Air quality prediction is fundamental to urban planning (Lai et al., 2023; Liu et al., 2025). Existing methodologies generally fall into two distinct paradigms:

Physics-Based Branch. These approaches construct ODEs/PDEs to model mechanisms like diffusion and advection that describe the spatial transport of atmospheric pollutants (Li et al., 2023; Sokhi et al., 2021), solved via numerical techniques (Özişik et al., 2017; Ŝolín, 2005). However, these models face two major limitations: (1) idealized closed-system assumptions that neglect open-boundary dynamics, and (2) prohibitive computational costs at large spatial scales (Lai et al., 2023).

Deep Learning Branch. Conversely, data-driven approaches leverage historical and auxiliary data for spatiotemporal modeling of evolving pollutant concentrations across distributed monitoring stations (Han et al., 2021, 2023). Beyond early shallow methods (Sánchez et al., 2013; Su et al., 2023), deep architectures—including RNNs, Transformers, and GNNs—excel at capturing complex non-linear

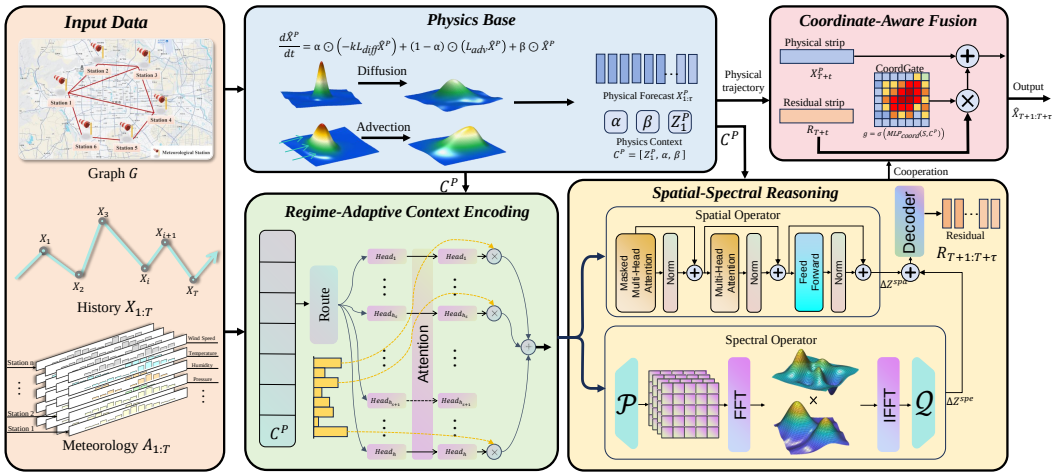


Figure 2: Overview of the proposed $\boldsymbol{PAS^2}$ Strategy. Physical branch establishes a deterministic baseline and conditions the neural branch. Data Driven branch consists of a Regime-Adaptive Context Encoder for dynamic state initialization and a $\boldsymbol{S^2}$ Reasoner that synergizes graph-based spatial operators with frequency-domain spectral operators to capture fine-grained residual fluctuations. A coordinate-aware fusion module combines the physical anchor and the results from neural calculation.

dependencies (Yu et al., 2018; Luo et al., 2019; Liang et al., 2023; Zhou et al., 2024; Liu et al., 2024). Yet, these "black-box" models often overlook essential physical laws, leading to confused correlations and inconsistencies under distribution shifts (Karpatne et al., 2017) (see Appendix C for details).

2.2 Physics-Guided Network Architecture

To bridge the gap between physical consistency and data-driven expressivity, recent studies have explored Physics-Guided Learning, which generally falls into two categories according to how physical knowledge is incorporated into the learning process.

Constraint-Based Approaches. The most direct strategy imposes physical laws as regularization terms in the loss function (Raissi et al., 2019; Shi et al., 2021). By penalizing deviations from governing PDEs, these models encourage physically consistent patterns and improve generalizability (Greydanus et al., 2019; Choi et al., 2023). However, this paradigm relies heavily on precise physical knowledge; strictly enforcing approximate equations on noisy data can introduce strong inductive biases that limit learning capacity or lead to optimization pathologies (Karpatne et al., 2024).

Hybrid Architectural Approaches. Currently, a more structural attempt is to integrate scientific knowledge deeply into network components rather than relying solely on loss-level constraints, typically by embedding differentiable solvers or Neural ODEs (Chen et al., 2018; Hwang et al., 2021) to parameterize transport dynamics. Recent works in air quality, such as AirPhyNet (Hettige et al., 2024) and others (Ji et al., 2022; Wang et al., 2022), have successfully adopted this hybrid route. Despite their progress, these methods often struggle to align deep latent representations with explicit physical quantities (Nascimento et al., 2021; Verma et al., 2024). Furthermore, most alternatives employ a parallel coupling design, where physical and neural branches operate concurrently. Such coupling may lead to functional redundancy, as the neural branch can dominate optimization and marginalize the physical prior, failing to achieve true synergy. In contrast, PAS^2 uses an independently supervised physical predictor as both a detached anchor and a conditioning signal for residual learning.

3 Methodology

Unlike conventional hybrid paradigms that entangle predictors, we enforce a strict separation of concerns, reformulating forecasting as a dual-process problem. A physical estimator establishes deterministic macroscopic trends and conditions residual reasoning, while a neural reasoner is

dedicated to capturing complex non-linear residuals. As shown in Fig. 2, we realize this Physics-Anchored Residual Correction scheme through three strategic stages that organize physical estimation and subsequent residual reasoning: (i) Physics-Anchored Decomposition (PAD), (ii) Regime-Adaptive Context Encoding (RACE), and (iii) Spatial-Spectral (S^2) Reasoning. Formally, we define the air quality network as a geospatial graph $G = (\mathcal{V}, \mathcal{E})$ with N monitoring stations. Given historical pollutant concentrations $X_{1:T} \in \mathbb{R}^{T\times N\times 1}$ and meteorological covariates $A_{1:T} \in \mathbb{R}^{T\times N\times D}$, our objective is to learn a mapping to predict future concentrations $X_{T+1:T+\tau} \in \mathbb{R}^{\tau\times N\times 1}$ over a horizon τ.

3.1 Physics-Anchored Decomposition

To reduce functional redundancy and direct optimization interference in parallel hybrid architectures, we formulate the forecasting task as a Physics-Anchored Residual Learning problem. We postulate that for each prediction step $t \in \{1, \dots, \tau\}$, the observed atmospheric state X_{T+t} is a superposition of a macroscopic physical baseline $\hat{X}^P_{T+t}$ and a complementary learnable residual $\hat{X}^R_{T+t}$.

To explicitly capture the distinct input dependencies and parameter spaces, we abstractly formulate two components via bipartite functional mappings:

$$X_{T+t} = \underbrace{\Phi(X_T, A_T, \hat{A}_{T+1:T+t}, \Theta_{\text{phy}})}_{\hat{X}^P_{T+t}} + \underbrace{\Psi(X_{1:T}, \hat{X}^P_{1:T+\tau}, \Theta_{\text{net}})}_{\hat{X}^R_{T+t}}, \tag{1}$$

where $\hat{A}_{T+1:T+t}$ denotes future meteorological covariates predicted from historical observations $A_{1:T}$. The physical operator $\Phi(\cdot)$ uses these predictions and the latest observed states X_T and A_T to establish the physical baseline. Conversely, the neural operator $\Psi(\cdot)$ estimates the fine-grained residual by conditioning on both the historical sequence $X_{1:T}$ and the derived physical trajectory $\hat{X}^P_{1:T+\tau}$, parameterized by network weights Θ_{net}.

3.1.1 Physical Micro-Operator Construction

To ground the learnable dynamics in fluid mechanics, we specify the diffusion and advection operator forms while learning the parameters that control their contributions. We consider the monitoring network as a Riemannian manifold approximated by a graph, where the transport is governed by two distinct mechanisms: isotropic diffusion and anisotropic advection.

Isotropic Diffusion Operator (L_{diff}). Representing the gradient-driven dispersion (Fick's Law), diffusion is naturally undirected and dependent solely on spatial proximity. We construct the diffusion Laplacian $L_{\text{diff}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}^{\text{diff}}$, where the adjacency weight $\mathbf{A}^{\text{diff}}_{ij}$ is defined by a radial basis function (RBF) kernel with a learnable bandwidth σ:

$$\mathbf{A}^{\text{diff}}_{ij} = \exp\left(-\frac{\text{dist}(\mathbf{p}_i, \mathbf{p}_j)^2}{\sigma^2}\right) \cdot \mathbb{I}(\text{dist}_{ij} < \epsilon), \tag{2}$$

where $\mathbf{p}$ denotes geocoordinates and $\mathbb{I}(\cdot)$ enforces sparse locality. This operator ensures smooth mass transfer solely driven by concentration gradients.

Anisotropic Advection Operator (L_{adv}). Driven by the wind field $\mathbf{v}(t)$, we employ an Upwind Discretization Scheme to enforce directional causality. The dynamic weight $\mathbf{A}^{\text{adv}}_{ji}$ is derived by projecting $\mathbf{v}_j$ onto the edge manifold $\mathbf{e}_{ji} = \mathbf{p}_i - \mathbf{p}_j$:

$$\mathbf{A}^{\text{adv}}_{ji}(t) = \frac{\langle \mathbf{v}_j(t), \mathbf{e}_{ji} \rangle_+}{\|\mathbf{e}_{ji}\|_2} = \max\left(0, \|\mathbf{v}_j\| \cos\theta_{\mathbf{v},\mathbf{e}}\right). \tag{3}$$

Here, $\langle\cdot\rangle_+$ isolates downwind flows, and L_{adv} calculates the net flux divergence: $[L_{\text{adv}}X]_i = \sum_{j\in\mathcal{N}(i)} \mathbf{A}^{\text{adv}}_{ji} X_j - X_i \sum_k \mathbf{A}^{\text{adv}}_{ik}$.

3.1.2 Physical Anchor: From Closed-Loop to Open-Boundary

Building upon these rigorous micro-operators, we formally examine the global properties of the transport system. Note that standard graph-based diffusion-advection equations implicitly assume a closed system. We formalize this limitation through the following theorem.

Theorem 3.1 (Conservation of Mass in Closed Systems)**.** *Given a closed geospatial graph G with the Laplacian operators L_{diff} and L_{adv} defined above, the total pollutant mass remains invariant over time under standard dynamics:*

$$\frac{d}{dt}\sum_{i=1}^{N} X_i(t) = \sum_{i=1}^{N}\left(-k[L_{\text{diff}}X]_i + [L_{\text{adv}}X]_i\right) = 0. \tag{4}$$

The mass leaving node i exactly equals the influx entering neighbors $\mathcal{N}(i)$. See Appendix B.2 for an idealized closed-system reference analysis.

While Theorem 3.1 holds mathematically, real-world urban environments are open systems characterized by continuous mass exchange (e.g., industrial emissions as sources, deposition as sinks) that violates the zero-sum assumption. To bridge this gap, a Boundary-Aware formulation (Tian et al., 2025) is employed. Practically, we introduce a learnable source-sink term $\beta \odot \hat{X}^P$ to explicitly break the conservation constraint. The governing equation for our Physical Anchor $\hat{X}^P$ is thus formulated as:

$$\frac{d\hat{X}^P}{dt} = \alpha \odot \left(-kL_{\text{diff}}\hat{X}^P\right) + (1-\alpha) \odot \left(L_{\text{adv}}\hat{X}^P\right) + \beta \odot \hat{X}^P. \tag{5}$$

Here, $\alpha \in [0,1]^N$ is a learnable mixing gate that modulates the local dominance between diffusion and advection, while $\beta \in \mathbb{R}^N$ captures the effective net contribution of unobserved external sources (generation, $\beta_i > 0$) or sinks (dissipation, $\beta_i < 0$). The physical baseline trajectory $\hat{X}^P$ is then generated by integrating Eq. (5) starting from X_T via a numerical ODE solver over the specified prediction horizon, capturing the macroscopic trends to anchor the subsequent neural reasoning.

3.1.3 Physical Context Projection

To effectively condition the neural branch on this physical regime without overwriting the constraints, we construct a station-wise Physical Context vector C_i^P. This context encapsulates the latent projection of the physical trajectory over the prediction horizon and the explicit dynamic parameters learned by the estimator for each monitoring station:

$$C_i^P = \left[\text{Enc}_{\text{phy}}(\hat{X}_{:,i}^P), \alpha_i, \beta_i\right]. \tag{6}$$

Here, $\alpha_i \in [0,1]$ and $\beta_i \in \mathbb{R}$ correspond to the i-th scalar components of the learned physical parameter vectors α and β from Eq. (5), representing the site-specific transport characteristics for station i. This context C^P acts as a physics-aware prompt for subsequent residual reasoning.

3.2 Regime-Adaptive Context Encoding

To capture the shifting dependencies governed by heterogeneous factors, RACE functions as a dynamic mechanism selector. First, temporal representations $H_{1:T}$ are extracted via a GRU from the augmented input $\tilde{X}_{t,i} = \text{Concat}(X_{t,i}, A_{t,i}, C_i^P)$. Then, a Physics-Anchored Top-m Routing is proposed. We conceptualize K attention heads as Latent Environmental Experts $\{\mathcal{A}_k\}$. The routing scores are computed as $s_{k,i} = \mathbf{w}_k^\top [H_{T,i}\,;\, C_i^P] + b_k$, explicitly conditioned on the physical context. This allows the expert selection to adapt to different temporal histories and inferred physical regimes across monitoring stations. Sparsity is enforced by selecting the top-m experts ($\mathcal{K}_i$) and computing mixing weights $\pi_{k,i}$ via a sparse softmax. The final physics-aware initial state is aggregated as $Z^R(0) = \text{Proj}(\sum_{k\in\mathcal{K}_i} \pi_{k,i}\mathcal{A}_k(H_{:,i}))$, ensuring the neural branch initialization aligns with the underlying transport regime.

3.3 Spatial-Spectral (S^2) Reasoning

The core challenge in modeling the fine-grained residual Ψ lies in the spatial-spectral distinctiveness of pollution dynamics: while background diffusion tends to produce smooth spatial patterns, abrupt pollution changes can create localized spatial contrasts involving a broad range of frequency components. Standard spatial aggregators tend to emphasize local smoothness, potentially oversmoothing these sharp contrasts. In our framework, the Fourier transform operates on the spatial grid rather than the temporal axis, enabling global interactions among spatial frequency modes to complement the temporal representations learned by RACE.

To resolve this limitation, we introduce the S^2 Reasoner, which evolves the latent residual state $Z^R(s)$ via a Neural ODE defined on the interval $s \in (0, 1]$. The derivative function $\mathcal{F}_{S^2}$ synergizes two complementary pathways to disentangle stable trends from volatile extremes:

$$\frac{\mathrm{d}Z^R(s)}{\mathrm{d}s} = \mathcal{F}_{S^2}\big(Z^R(s); G, C^P\big). \tag{7}$$

3.3.1 Spatial Operator

The spatial pathway focuses on maintaining local consistency when propagating pollutant representations across neighboring monitoring stations. However, standard graph convolutions often ignore topographic barriers (e.g., mountains blocking diffusion) that may affect the effective connectivity between geographically close stations. To address this, we modulate the adjacency matrix with a rigorous geophysical constraint. We define a topographic connectivity mask $\mathcal{M}_{ij}$ using the supremum of elevation along the path:

$$\mathcal{M}_{ij} = \mathbb{I}\left(\sup_{\lambda\in[0,1]} E\big(\lambda\rho_i + (1-\lambda)\rho_j\big) < \tau_h\right), \tag{8}$$

where $E(\cdot)$ is the elevation function at coordinate ρ, and τ_h is the barrier threshold. We then employ a Physical-Context Modulated GAT constrained by this topology:

$$\Delta Z_i^{\mathrm{spa}} = \sum_{j\in\mathcal{N}(i)} \underbrace{\mathcal{M}_{ij}\cdot a_{ij}}_{\text{Topo-Aware Attn}} \mathbf{W} Z_j^R(s). \tag{9}$$

3.3.2 Spectral Operator

To overcome the limited receptive field of spatial operators and explicitly model global teleconnections, we introduce a Frequency-Domain Transformer Pathway. This module combines frequency-domain global mixing with Transformer sequence modeling to capture long-range dependencies efficiently. Our Fourier-based attention models global interactions among spatial frequency modes, complementing the local transport structure preserved by the Spatial Operator.

Let $\mathbf{U} \in \mathbb{R}^{H\times W\times d}$ be the grid-projected features. We normalize station coordinates within each dataset's fixed geographical bounding box and interpolate station features onto a 32×32 grid using four-nearest-neighbor inverse-distance weighting. The projection depends on geographical coordinates rather than station ordering, with interpolation weights and boundary treatment detailed in Appendix T.1. The spectral evolution follows a Transform-Process-Invert paradigm:

Spectral Transformation. We first convert the spatial features $\mathbf{U}$ into the frequency domain using 2D FFT ($\mathcal{F}$), mapping the spatial grid to its spectral representation $\widehat{\mathbf{U}}$:

$$\widehat{\mathbf{U}} = \mathcal{F}(\mathbf{U}) \in \mathbb{C}^{H\times W\times d}. \tag{10}$$

In this domain, global spatial patterns are compacted into local spectral peaks, allowing the subsequent network to process global information efficiently.

Frequency-Domain Transformer Processing. To model the complex interactions between different frequency modes (which correspond to different spatial scales), we treat the spectrum as a sequence. The spectral features are flattened into a sequence of tokens $\widehat{\mathbf{S}} \in \mathbb{C}^{H\times W\times d}$, followed by the processing of a Frequency Transformer Encoder, which consists of stacked layers of Multi-Head Self-Attention (MSA) and Feed-Forward Networks (FFN). To handle complex-valued input, we concatenate the real and imaginary parts as feature channels. The update rule for a single layer is:

$$\begin{aligned}
\widehat{\mathbf{S}}' &= \mathrm{LayerNorm}\Big(\widehat{\mathbf{S}} + \mathrm{MSA}(\widehat{\mathbf{S}})\Big),\\
\widehat{\mathbf{S}}'' &= \mathrm{LayerNorm}\Big(\widehat{\mathbf{S}}' + \mathrm{FFN}(\widehat{\mathbf{S}}')\Big).
\end{aligned} \tag{11}$$

Here, $\mathrm{MSA}(\cdot)$ allows the model to dynamically attend to significant frequency components (e.g., dominant periodicities or high-frequency transients) regardless of their spectral position, while $\mathrm{FFN}(\cdot)$ processes the features point-wise.

Inverse Spectral Transformation. The refined spectral tokens $\widehat{\mathbf{S}}''$, which now encode globally contextualized information, are reshaped back to the spectral grid $\widehat{\mathbf{U}}'$ and transformed back to the spatial domain via the Inverse FFT:

$$\Delta Z^{\mathrm{spe}} = Q_\theta\Big(\mathcal{F}^{-1}(\widehat{\mathbf{U}}')\Big). \tag{12}$$

This operation distributes global context learned in the frequency domain back to local spatial coordinates, effectively recovering fine-grained details and teleconnections.

3.3.3 Dual-Domain Synergization

To dynamically balance local diffusion modeled by spatial branch and global teleconnections captured by frequency branch, we employ a context-aware gating mechanism to adapt spectral corrections across different monitoring stations. The final derivative of the residual state is defined as:

$$\frac{\mathrm{d}Z^R}{\mathrm{d}s} = \Delta Z^{\mathrm{spa}} + \sigma\Big(\mathcal{G}(Z^R, C^P)\Big) \odot \Delta Z^{\mathrm{spe}}, \tag{13}$$

where $\mathcal{G}(\cdot)$ is a lightweight neural network that computes a spatially-adaptive gate $\mathbf{g} \in [0,1]^N$ conditioned on the physical context C^P. This selectively applies spectral corrections to regions with non-local dynamics or rapid regime shifts.

3.4 Learning Objective and Optimization Separation

PAS^2 is trained to separate the roles of the physical estimator and the neural residual branch, rather than optimizing them as competing predictors. Let Θ_{phy} and Θ_{net} denote the parameters of the physical estimator and neural branch, respectively. The physical trajectory and context condition the residual branch through stop-gradient:

$$\bar{X}^P = \mathrm{sg}(\hat{X}^P), \qquad \bar{C}^P = \mathrm{sg}(C^P), \tag{14}$$

where $\mathrm{sg}(\cdot)$ blocks gradients to the physical branch. The final prediction is

$$\hat{X}_{T+t} = \bar{X}^P_{T+t} + \hat{X}^R_{T+t}, \qquad \hat{X}^R_{T+t} = \Psi(X_{1:T}, \bar{X}^P_{1:T+\tau}, \bar{C}^P; \Theta_{\mathrm{net}}). \tag{15}$$

Forecasting Loss. The residual branch is optimized by minimizing the anchored prediction error over all forecasting steps and monitoring stations:

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{\tau N}\sum_{t=1}^{\tau}\sum_{i=1}^{N}\Big\|\bar{X}^P_{T+t,i} + \hat{X}^R_{T+t,i} - X_{T+t,i}\Big\|_2^2. \tag{16}$$

Physics Supervision Loss. The physical anchor is independently supervised to preserve macroscopic transport dynamics:

$$\mathcal{L}_{\mathrm{phy}} = \frac{1}{\tau N}\sum_{t=1}^{\tau}\sum_{i=1}^{N}\Big\|\hat{X}^P_{T+t,i} - X_{T+t,i}\Big\|_2^2. \tag{17}$$

Open-Boundary Regularization. To avoid implausible source–sink compensation, we regularize the open-boundary source–sink coefficients:

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{N}\sum_{i=1}^{N}\|\beta_i\|_2^2. \tag{18}$$

The overall objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{pred}} + \lambda_1\mathcal{L}_{\mathrm{phy}} + \lambda_2\mathcal{L}_{\mathrm{reg}}, \tag{19}$$

where λ_1 and λ_2 control the strength of physical supervision and regularization.

Although written as a unified objective, the stop-gradient operation ensures that $\mathcal{L}_{\mathrm{pred}}$ updates only the neural residual branch, while $\mathcal{L}_{\mathrm{phy}}$ and $\mathcal{L}_{\mathrm{reg}}$ update only the physical estimator. Thus, the forecasting and physics-related gradients lie in disjoint parameter subspaces:

$$\langle \nabla_\Theta \mathcal{L}_{\mathrm{pred}}, \nabla_\Theta\left(\lambda_1\mathcal{L}_{\mathrm{phy}} + \lambda_2\mathcal{L}_{\mathrm{reg}}\right)\rangle = 0. \tag{20}$$

This structural separation removes direct cross-branch gradient interference.

4 Experiments and Analysis

4.1 Experimental Setting

Datasets and Setup. We evaluate PAS^2 on two standard benchmarks: the city-level *Beijing* and national-level *KnowAir* datasets. Following recent literature (Hettige et al., 2024; Wang et al., 2020), we forecast $PM_{2.5}$ conditioned on five auxiliary meteorological covariates (Tian et al., 2025). We adopt a 3-hour resolution, utilizing 72 hours (24 steps) of history to predict the subsequent 72 hours. Furthermore, we test model robustness on *Sudden Change* scenarios (Liang et al., 2023), defined as instances where $PM_{2.5}$ exceeds 75 $\mu g/m^3$ and fluctuates by over $\pm 20\,\mu g/m^3$ within a 3-hour window. Detailed dataset statistics are provided in Appendix F.

Baselines. We benchmark our framework against recent methods spanning four categories: 1) *Classical methods*, including HA and VAR; 2) *Continuous-time models* based on Neural ODEs, specifically Latent-ODE (Rubanova et al., 2019), ODE-RNN, and ODE-LSTM (Lechner & Hasani, 2020); 3) *Spatio-temporal deep learning models*, encompassing both generic graph/attention-based networks (DCRNN (Li et al., 2018b), STGCN (Yu et al., 2018), ASTGCN (Guo et al., 2019), GTS (Shang & Chen, 2021), GMAN, STTN) and domain-specific air quality models ($PM_{2.5}$-GNN (Wang et al., 2020), AirFormer (Liang et al., 2023)); and 4) *Physics-guided models*, namely AirPhyNet (Hettige et al., 2024) and Air-DualODE (Tian et al., 2025).

Evaluation metrics. We evaluate performance using three standard regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE), where lower values indicate better accuracy.

4.2 Performance Comparison

Overall Performance. Table 1 compares PAS^2 with classical, neural ODE, spatio-temporal, and physics-guided baselines. Deep models substantially outperform HA and VAR, while domain-specific air-quality models generally improve over generic spatio-temporal predictors, confirming the value of atmospheric inductive biases. Among physics-guided methods, PAS^2 consistently surpasses AirPhyNet and Air-DualODE, supporting the effectiveness of using the physical estimator as an anchor rather than as a peer predictor. The gains are especially pronounced on *Sudden Change* subsets, where S^2 Reasoning captures high-frequency residuals and RACE adapts to rapid regime shifts. A more detailed baseline-level discussion is provided in Appendix J.

Table 1: Overall comparisons of prediction performance. The best results are highlighted in red and the second-best in purple.

Methods	KnowAir						Beijing					
	3days			Sudden Change			3days			Sudden Change		
	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
HA	29.40	39.18	0.61	58.05	81.29	0.57	58.88	87.51	0.82	180.02	245.79	0.86
VAR	27.66	36.56	0.58	56.26	75.10	0.55	68.12	119.95	1.15	206.20	286.45	1.18
LatentODE	23.01	32.73	0.51	52.99	72.48	0.53	38.43	84.02	0.57	197.56	266.50	1.04
ODE-RNN	22.59	32.48	0.50	51.71	71.54	0.52	43.60	80.33	0.67	179.57	247.82	0.85
ODE-LSTM	23.80	33.82	0.49	53.83	73.06	0.54	61.87	90.64	0.82	173.88	232.56	0.85
DCRNN	23.54	35.41	0.53	61.12	84.28	0.63	44.48	82.92	0.77	177.95	247.09	0.90
STGCN	24.17	34.95	0.51	59.58	78.70	0.62	36.81	73.35	0.58	161.03	226.49	0.77
ASTGCN	19.21	30.73	0.43	50.59	72.54	0.53	37.45	76.93	0.58	176.79	247.08	0.89
GTS	24.48	34.87	0.54	53.29	73.79	0.53	50.04	84.83	0.77	183.60	251.69	0.87
GMAN	20.30	32.07	0.45	54.68	75.62	0.56	57.62	91.30	0.79	192.48	259.90	0.97
STTN	26.41	38.67	0.54	58.44	76.57	0.56	64.48	89.06	0.78	160.22	222.35	0.79
$PM_{2.5}$-GNN	22.47	33.68	0.52	53.25	73.32	0.53	38.13	79.91	0.59	178.30	247.82	0.94
AirFormer	18.89	29.83	0.44	51.15	71.26	0.52	33.97	70.23	0.54	153.04	221.70	0.73
AirPhyNet	20.52	32.03	0.45	54.56	75.69	0.55	32.13	71.81	0.53	152.14	220.72	0.74
Air-DualODE	19.06	29.79	0.43	51.32	72.85	0.51	38.35	82.71	0.58	156.50	229.96	0.79
PAS^2	18.70	29.53	0.42	48.41	68.01	0.48	31.24	68.29	0.48	148.46	217.80	0.72

4.3 Ablation Study

Effect of Physics-Anchored Decomposition. As shown in Table 2, our framework outperforms three structural variants: *w/o Physics Base*, *w/o Residual Anchoring*, and *Parallel Coupling*. Notably, the failure of *Parallel Coupling* highlights the *functional redundancy* and *gradient competition* inherent in peer-predictor designs. By strictly dedicating the neural branch to residual modeling, our physics-anchored strategy mitigates these conflicts and efficiently captures non-stationary dynamics beyond the physical baseline. Further comparisons with a physics-only predictor and a parameter-matched learned anchor are provided in Appendix P. The results demonstrate the benefit of residual correction and the prescribed physical transport operators, respectively.

Efficacy of S^2 Reasoning. We validate the S^2 Reasoner's capacity to resolve the "Low-Pass Dilemma" via component ablation (Table 2, Fig. 3). The sharp error increase in *w/o Spectral Path* on the *Sudden Change* subset confirms that spectral features recover high-frequency transients typically lost to spatial oversmoothing. Conversely, *w/o Spatial Path* degrades overall accuracy, underscoring the need for smooth diffusion modeling. Finally, the failure of *w/o Dual-Domain Gating* emphasizes the necessity of adaptively balancing spatial smoothness and spectral sharpness. Complementary experiments in Appendix P replace the Spectral Transformer with parameter-matched Global Attention and Graph Transformer modules. The full model achieves lower errors on both datasets, particularly for sudden changes.

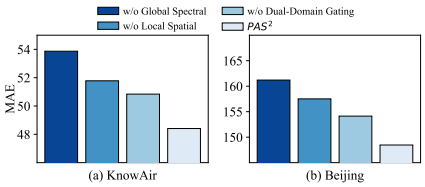


Figure 3: Ablation of S^2 Reasoning components on sudden-change subsets.

Table 2: Ablation study of PAS^2 on KnowAir and Beijing datasets. The best results (full model) are highlighted in red and the second-best in purple.

Variants	KnowAir						Beijing					
	3days			Sudden Change			3days			Sudden Change		
	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
w/o Physics Base	22.15	34.04	0.51	56.82	82.04	0.59	42.11	81.55	0.65	173.23	241.28	0.86
w/o Residual Anchoring	20.44	31.88	0.46	53.15	74.19	0.54	35.60	76.35	0.55	162.36	230.55	0.80
Parallel Coupling	21.07	32.48	0.47	54.76	75.93	0.55	36.85	77.21	0.57	165.42	234.05	0.82
w/o Spectral Path	20.74	31.96	0.46	53.87	76.82	0.56	34.88	75.51	0.54	161.20	233.83	0.81
w/o Spatial Path	19.92	31.18	0.45	51.78	72.05	0.52	33.53	72.83	0.52	157.51	227.36	0.78
w/o Dual-Domain Gating	19.25	30.43	0.44	50.79	70.81	0.51	32.27	70.37	0.51	154.13	223.29	0.75
w/o MoHAttention	19.85	30.92	0.44	53.58	73.71	0.53	33.45	72.16	0.52	159.82	228.17	0.79
w/o Dynamic Gating	19.33	30.25	0.43	51.47	71.38	0.51	32.02	69.88	0.49	154.67	225.10	0.76
w/o Expert Sparsity	19.12	30.18	0.43	50.84	70.47	0.50	31.92	69.57	0.49	152.64	222.48	0.76
w/o Phy. Routing	19.41	30.65	0.44	52.93	73.15	0.53	32.81	71.24	0.51	158.38	228.61	0.78
w/o Phy. Injection	20.27	31.63	0.45	53.04	73.98	0.53	34.17	73.10	0.53	160.54	231.26	0.79
w/o Phy. Encoder	19.63	30.67	0.44	51.24	71.61	0.51	32.60	70.68	0.51	153.82	224.35	0.77
PAS^2	18.70	29.53	0.42	48.41	68.01	0.48	31.24	68.29	0.48	148.46	217.80	0.72

Effect of Regime-Adaptive Encoder. To isolate the benefits of adaptive, physics-aware regime selection, we evaluate four variants: *w/o MoHAttention*, *w/o Dynamic Gating*, *w/o Expert Sparsity*, and *w/o Physics-Conditioned Routing*. Fig. 4 exhibits performance degrades consistently across all settings, with significantly sharper drops in *sudden-change* scenarios. This corroborates that pollutant transport is governed by distinct, evolving modes. Crucially, it demonstrates that physics-conditioned routing is essential for dynamically selecting the optimal "regime expert" to resolve non-stationary dependencies.

Necessity of Physical Context. Ablating physics-aware conditioning (*w/o Phy. Injection* and *w/o Phy. Encoder"*) leads to marked performance degradation (Table 2), verifying that residual dynamics are inextricably linked to physical states. Explicit context injection effectively constrains the solution space, enforcing physics-consistent corrections rather than unconstrained mappings (see detailed analysis in Appendix P).

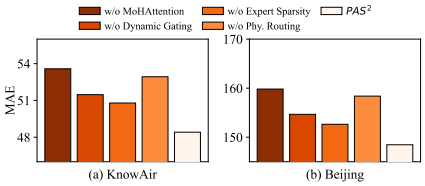


Figure 4: Ablation of RACE.

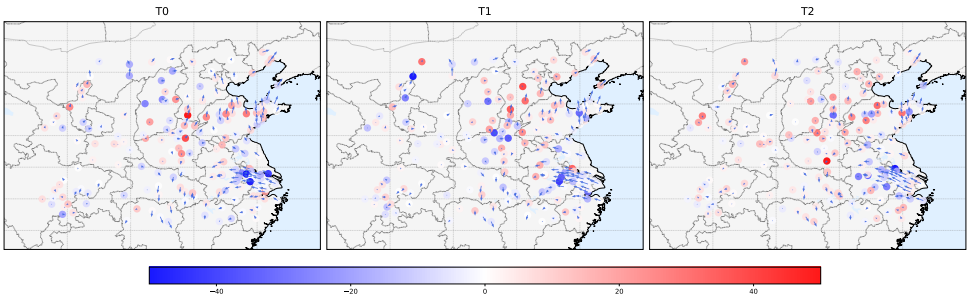


Figure 5: Visualization of wind-driven residual dynamics. Blue arrows denote wind vectors; red/blue bubbles indicate $PM_{2.5}$ accumulation/dissipation.

4.4 Case Study

We visualize $PM_{2.5}$ residual evolution to examine how PAS^2 captures real-world atmospheric dynamics. Fig. 5 gives a typical pollutant transport in Eastern China over 3 consecutive timestamps (T_0, T_1, T_2), overlaying the predicted concentration gradients with the corresponding wind vector fields.

Capturing Deterministic Advection. A macroscopic physical consistency is immediately observable. At T_0, the region is dominated by a strong Northwest wind field. As the system evolves to T_1 and T_2, we observe a coherent "Upwind Dissipation, Downwind Accumulation" pattern. Specifically, stations in the upwind northern regions exhibit deep blue bubbles, indicating a rapid depletion of pollutants due to strong ventilation. Conversely, the downwind southern regions are marked by dense red bubbles, signifying the accumulation of the transported mass. This precise alignment between the pollution trajectory and the wind vectors corroborates that our PAD successfully anchors the neural predictions to the deterministic transport laws (Eq. 5). Meanwhile, the three snapshots reveal heterogeneous local changes superimposed on this prevailing transport pattern. The magnitude of accumulation and dissipation varies across monitoring stations, rather than forming a uniformly evolving concentration field. These spatial contrasts illustrate how coherent regional transport can coexist with fine-grained pollutant fluctuations, providing a natural connection between the physical anchor and the complementary residual reasoning in PAS^2.

Resolving High-Frequency Transients. Beyond the directional trend, the visualization highlights the critical role of the S^2 Reasoner. Standard spatial models, acting as low-pass filters, typically oversmooth such rapid regime shifts, resulting in underestimated gradients (i.e., pale or neutral bubbles). In contrast, our model predicts highly saturated red and blue zones, sharply distinguishing regions of drastic change from the stable background. Across successive timestamps, the spatial extent and intensity of these contrasting regions continue to evolve, while the most pronounced changes remain distinguishable from the relatively stable surrounding areas. Such evolving patterns highlight the importance of preserving sharp local variations together with the broader spatial distribution of pollutant changes. This indicates that the Spectral Operator successfully recovers the high-energy components of the residual signal in the frequency domain. By explicitly modeling these "drastic fluctuations" rather than smoothing them out, the framework accurately captures the fine-grained transient dynamics that are critical for early warning in high-volatility scenarios.

5 Conclusion and Future Works

In this paper, we introduced PAS^2 Strategy to address the data homogeneity and "low-pass dilemma" in air quality forecasting. An independently supervised physical anchor models transport trends and conditions residual reasoning, while the S^2 Reasoner captures complementary broadband fluctuations. Extensive experiments on real-world datasets demonstrate that our method significantly outperforms state-of-the-art baselines, exhibiting superior robustness in capturing abrupt regime shifts and extreme peaks. However, our current physical formulation simplifies complex chemical transformations (e.g., secondary aerosol formation) into implicit residual parameters. In the future, we plan to augment the physical anchor with explicit chemical kinetics to enhance interpretability and further extend the PAS^2 paradigm to general spatiotemporal forecasting tasks characterized by volatile dynamics.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62676366 and 62276232, and the Natural Science Foundation of Zhejiang Province under Grant No. LZ24F030012.

References

Azimi, M. N. and Rahman, M. M. Unveiling the health consequences of air pollution in the world's most polluted nations. *Scientific Reports*, 14(1):9856, 2024.

Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., Tong, Y., Xu, B., Bai, J., Tong, J., et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems*, 33:17766–17778, 2020.

Chapman, A. *Semi-Autonomous Networks: Effective Control of Networked Systems through Protocols, Design, and Modeling*. Springer, 2015.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31:6572–6583, 2018.

Cheng, L., Lu, S., and Shi, R. Research on environmental pollution caused by acid rain. *Highlights Sci Eng Technol*, 91:25–29, 2024.

Choi, J., Hong, S., Park, N., and Cho, S.-B. Gread: Graph neural reaction-diffusion networks. In *International Conference on Machine Learning*, pp. 5722–5747. PMLR, 2023.

Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, 32:15353–15363, 2019.

Guen, V. L. and Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11484, 2020.

Guo, S., Lin, Y., Feng, N., Song, C., and Wan, H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 922–929, 2019.

Han, J., Liu, H., Zhu, H., Xiong, H., and Dou, D. Joint air quality and weather prediction based on multi-adversarial spatiotemporal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4081–4089, 2021.

Han, J., Zhang, W., Liu, H., and Xiong, H. Machine learning for urban air quality analytics: A survey. *arXiv preprint arXiv:2310.09620*, 2023.

Hettige, K. H., Ji, J., Xiang, S., Long, C., Cong, G., and Wang, J. Airphynet: Harnessing physics-guided neural networks for air quality prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

Hwang, J., Choi, J., Choi, H., Lee, K., Lee, D., and Park, N. Climate modeling with neural diffusion equations. In *2021 IEEE international conference on data mining (ICDM)*, pp. 230–239. IEEE, 2021.

Hwang, Y. and Lim, D.-Y. Dual cone gradient descent for training physics-informed neural networks. *Advances in Neural Information Processing Systems*, 37:98563–98595, 2024.

Ji, J., Wang, J., Jiang, Z., Jiang, J., and Zhang, H. Stden: Towards physics-guided neural networks for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4048–4056, 2022.

Jiang, J., Han, C., Zhao, W. X., and Wang, J. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4365–4373, 2023a.

Jiang, R., Wang, Z., Yong, J., Jeph, P., Chen, Q., Kobayashi, Y., Song, X., Fukushima, S., and Suzumura, T. Spatio-temporal meta-graph learning for traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 8078–8086, 2023b.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

Karpatne, A., Jia, X., and Kumar, V. Knowledge-guided machine learning: Current trends and future prospects. *arXiv preprint arXiv:2403.15989*, 2024.

Lai, S., Xu, Z., Zhang, W., Liu, H., and Xiong, H. Large language models as traffic signal control agents: Capacity and opportunity. *CoRR*, 2023.

Lechner, M. and Hasani, R. Learning long-term dependencies in irregularly-sampled time series. *arXiv preprint arXiv:2006.04418*, 2020.

Li, L., Khalili, R., Lurmann, F., Pavlovic, N., Wu, J., Xu, Y., Liu, Y., O'Sharkey, K., Ritz, B., Oman, L., et al. Physics-informed deep learning to reduce the bias in joint prediction of nitrogen oxides. *arXiv preprint arXiv:2308.07441*, 2023.

Li, Q., Han, Z., and Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3538–3545, 2018a.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018b.

Li, Z., Kovachki, N. B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A., et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., Zheng, Y., and Zimmermann, R. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14329–14337, 2023.

Liu, S., Shu, Y., Guo, C., and Yang, B. Learning generalizable skills from offline multi-task data for multi-agent cooperation. *arXiv preprint arXiv:2503.21200*, 2025.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Luo, Z., Huang, J., Hu, K., Li, X., and Zhang, P. Accuair: Winning solution to air quality prediction for kdd cup 2018. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1842–1850, 2019.

Nascimento, R. G., Corbetta, M., Kulkarni, C. S., and Viana, F. A. Hybrid physics-informed neural networks for lithium-ion battery modeling and prognosis. *Journal of Power Sources*, 513:230526, 2021.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Özişik, M. N., Orlande, H. R., Colaço, M. J., and Cotta, R. M. *Finite difference methods in heat transfer*. CRC press, 2017.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 32:5321–5331, 2019.

Sánchez, A. S., Nieto, P. G., Iglesias-Rodríguez, F. J., and Vilán, J. V. Nonlinear air quality modeling using support vector machines in gijón urban area (northern spain) at local scale. *International Journal of Nonlinear Sciences and Numerical Simulation*, 14(5):291–305, 2013.

Schiesser, W. E. *The numerical method of lines: integration of partial differential equations*. Elsevier, 2012.

Seo, S., Meng, C., and Liu, Y. Physics-aware difference graph networks for sparsely-observed dynamics. In *International Conference on Learning Representations*, 2019.

Shang, C. and Chen, J. Discrete graph structure learning for forecasting multiple time series. In *Proceedings of International Conference on Learning Representations*, 2021.

Shi, R., Mo, Z., and Di, X. 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.

Shi, R., Luo, J., Zhou, N., Liu, Y., Hong, C., Zhang, X.-P., and Chen, X. Phy-apmr: A physics-informed air pollution map reconstruction approach with mobile crowd-sensing for fine-grained measurement. *Building and Environment*, 272:112634, 2025.

Sokhi, R. S., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., Friedrich, R., Geels, C., Grönholm, T., Halenka, T., et al. Advances in air quality research–current and emerging challenges. *Atmospheric Chemistry and Physics Discussions*, 2021:1–133, 2021.

Ŝolín, P. *Partial differential equations and the finite element method*. John Wiley & Sons, 2005.

Su, M., Liu, H., Yu, C., and Duan, Z. A novel aqi forecasting method based on fusing temporal correlation forecasting with spatial correlation forecasting. *Atmospheric Pollution Research*, 14(4): 101717, 2023.

Tian, J., Liang, Y., Xu, R., Chen, P., Guo, C., Zhou, A., Pan, L., Rao, Z., and Yang, B. Air quality prediction with physics-guided dual neural odes in open systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

Verma, Y., Heinonen, M., and Garg, V. Climode: Climate and weather forecasting with physics-informed neural odes. In *The Twelfth International Conference on Learning Representations*, 2024.

Wang, J., Ji, J., Jiang, Z., and Sun, L. Traffic flow prediction based on spatiotemporal potential energy fields. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9073–9087, 2022.

Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., and Gao, F. 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*, pp. 163–166, 2020.

Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J., and Zhou, J. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024.

Wang, S., Bhartari, A. K., Li, B., and Perdikaris, P. Gradient alignment in physics-informed neural networks: A second-order optimization perspective. *arXiv preprint arXiv:2502.00604*, 2025.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Xu, M., Dai, W., Liu, C., Gao, X., Lin, W., Qi, G.-J., and Xiong, H. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020.

Yang, T., Qian, Z., Hang, N., and Liu, M. S-pinn: Stabilized physics-informed neural networks for alleviating barriers between multi-level co-optimization. *Computer Methods in Applied Mechanics and Engineering*, 447:118348, 2025.

Yi, K., Zhang, Q., Fan, W., Wang, S., Wang, P., He, H., An, N., Lian, D., Cao, L., and Niu, Z. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36:76656–76679, 2023.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 3634–3640. International Joint Conferences on Artificial Intelligence Organization, 2018.

Zheng, C., Fan, X., Wang, C., and Qi, J. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1234–1241, 2020.

Zhou, S., Wang, W., Zhu, L., Qiao, Q., and Kang, Y. Deep-learning architecture for pm2. 5 concentration prediction: A review. *Environmental Science and Ecotechnology*, 21:100400, 2024.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pp. 27268–27286. PMLR, 2022.

A Nomenclature and Notation Tables

We summarize the mathematical notations used throughout the paper, categorized by their functional roles within the PAS^2 framework.

Table 3: Problem Definition and Inputs

Symbol	Description
$G = (\mathcal{V}, \mathcal{E})$	Geospatial graph with N monitoring stations ($\mathcal{V}$) and edges ($\mathcal{E}$)
N	Number of monitoring stations
T, τ	Length of historical observation window and prediction horizon
$X_{1:T}$	Historical pollutant concentrations tensor, $X \in \mathbb{R}^{T \times N \times 1}$
$A_{1:T}$	Auxiliary meteorological covariates (e.g., wind, temp), $A \in \mathbb{R}^{T \times N \times D}$
$\hat{X}_{T+1:T+\tau}$	Predicted future pollutant concentrations
$\mathbf{p}_i$	Geocoordinates (latitude, longitude) of station i
$E(\cdot)$	Terrain elevation function for topographic constraints

Table 4: Physics-Anchored Decomposition & Micro-Operators

Symbol	Description
$\Phi(\cdot), \Psi(\cdot)$	Deterministic physical operator and learnable residual operator
$\hat{X}^P$	**Physical Anchor**: The deterministic baseline trajectory derived from ODEs
$\hat{X}^R$	**Residual Correction**: The complementary high-frequency dynamics
L_{diff}	Positive random-walk diffusion Laplacian constructed via RBF kernels
L_{adv}	Anisotropic Advection operator based on Upwind Discretization
$\mathbf{v}(t)$	Dynamic wind velocity vector field
$\mathbf{e}_{ji}$	Edge manifold vector projecting from station j to i
$\langle \cdot, \cdot \rangle_+$	Positive part operator used in the Upwind Scheme
α	Learnable mixing gate balancing diffusion vs. advection, $\alpha \in [0, 1]^N$
β	Source-sink coefficient for Open-Boundary correction, $\beta \in \mathbb{R}^N$
C_i^P	Physical Context vector encapsulating regime parameters for station i

Table 5: Spatial-Spectral Reasoning & Optimization

Symbol	Description
$\{\mathcal{A}_k\}_{k=1}^K$	Set of K Latent Environmental Experts in the RACE module
$s_{k,i}, \pi_{k,i}$	Relevance scores and sparse mixing weights for expert routing
$Z^R(s)$	Latent residual state evolving in continuous time $s \in [0, 1]$
$\mathcal{M}_{ij}$	Topographic connectivity mask blocking transport across barriers
$\mathcal{F}, \mathcal{F}^{-1}$	2D Fast Fourier Transform (FFT) and its inverse
$\widehat{\mathbf{U}}$	Spectral representation of spatial features in the frequency domain
$\Delta Z^{\text{spa/spe}}$	State updates from Local Spatial and Global Spectral operators
$\mathcal{G}(\cdot)$	Dual-domain gating mechanism balancing spatial/spectral paths
$\Theta_{\text{phy/net}}$	Learnable parameters for the physical estimator and neural networks
λ_1, λ_2	Hyperparameters for physics supervision and regularization weights

B Derivation of Physics-Anchored Transport Dynamics

To establish the physical anchor $\hat{X}^P$ in our PAS^2 framework, we ground our formulation in classical fluid mechanics. Let $C(\mathbf{x}, t)$ denote the spatiotemporal distribution of pollutant concentration. The transport mechanism is decomposed into two fundamental sub-processes: Diffusion and Advection.

Continuous Diffusion Equation: Diffusion is driven by the spatial gradient of pollutants, strictly moving mass from high-concentration to low-concentration regions. By Fick's Second Law, the continuous diffusion equation is expressed as:

$$\frac{\partial C}{\partial t} - k\nabla^2 C = 0, \tag{21}$$

where ∇^2 is the Laplace-Beltrami operator and k is the diffusion coefficient.

Continuous Advection Equation: Advection is driven by the dynamic meteorological wind field $\mathbf{v}(\mathbf{x}, t)$, propagating pollutants along the wind streamlines:

$$\frac{\partial C}{\partial t} + \nabla \cdot (C\mathbf{v}) = 0. \tag{22}$$

B.1 Spatial Discretization via Method of Lines

To solve Eq. (21) and Eq. (22) within our geospatial neural network, we employ the Method of Lines (MOL) (Schiesser, 2012), a standard technique in Computational Fluid Dynamics (CFD). MOL discretizes the continuous spatial PDEs into a grid of location-specific Ordinary Differential Equations (ODEs). Following the numerical principles in Chapman (2015) and the graph-based formulations adapted for air quality (Wang et al., 2020), we derive the nodal update rules.

For a monitoring station i with neighbors $\mathcal{N}(i)$, the **Discrete Diffusion Equation** is derived as:

$$\left.\frac{d\hat{X}_i^P}{dt}\right|_{\text{diff}} = k \sum_{j \in \mathcal{N}(i)} \mathbf{A}_{ij}^{\text{spa}} (\hat{X}_j^P - \hat{X}_i^P), \tag{23}$$

where $\mathbf{A}_{ij}^{\text{spa}}$ represents the static diffusion weight.

Simultaneously, the **Discrete Advection Equation** follows the upstream/downstream wind relations:

$$\left.\frac{d\hat{X}_i^P}{dt}\right|_{\text{adv}} = \underbrace{\sum_{j \in \mathcal{N}_{\text{in}}(i)} \mathbf{A}_{ji}^{\text{adv}} \hat{X}_j^P}_{\text{Wind Inflow}} - \underbrace{\sum_{k \in \mathcal{N}_{\text{out}}(i)} \mathbf{A}_{ik}^{\text{adv}} \hat{X}_i^P}_{\text{Wind Outflow}}, \tag{24}$$

where $\mathbf{A}_{ij}^{\text{adv}}$ captures the dynamic advection weight directed from i to j.

Aggregating these two discrete components via our learnable spatial gate $\alpha \in [0, 1]^N$, we obtain the governing base ODE for our anchor:

$$\frac{d\hat{X}_i^P}{dt} = \alpha_i \cdot \left.\frac{d\hat{X}_i^P}{dt}\right|_{\text{diff}} + (1 - \alpha_i) \cdot \left.\frac{d\hat{X}_i^P}{dt}\right|_{\text{adv}}. \tag{25}$$

B.2 Mass Conservation in an Idealized Closed-System Reference

We provide a formal conservation analysis for an idealized closed-system transport reference. This analysis is used to motivate the need for the open-boundary source–sink term in the physical anchor. It should not be interpreted as claiming that the full implemented physical anchor, which contains node-wise gates, row-normalized graph operators, and learnable source–sink effects, is exactly mass-conserving.

Consider an idealized closed graph with symmetric pairwise diffusion weights $w_{ij} = w_{ji} \geq 0$ and no external source or sink. The conservative diffusion dynamics are written in pairwise flux form:

$$\left.\frac{dx_i}{dt}\right|_{\text{diff}} = k \sum_{j \in \mathcal{N}(i)} w_{ij}(x_j - x_i). \tag{26}$$

The total mass variation induced by diffusion is

$$\begin{aligned}\frac{d}{dt}\sum_{i=1}^{N} x_i &= k\sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} w_{ij}(x_j - x_i) \\ &= \frac{k}{2}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} [w_{ij}(x_j - x_i) + w_{ji}(x_i - x_j)] \\ &= 0,\end{aligned} \tag{27}$$

where the second equality groups each unordered edge pair (i, j) and the last equality follows from $w_{ij} = w_{ji}$.

For advection, conservation can also be expressed through antisymmetric pairwise fluxes. Let q_{ij} denote the advective mass flux from node i to node j, satisfying

$$q_{ij} = -q_{ji}. \tag{28}$$

The conservative advection dynamics can be written as

$$\left.\frac{dx_i}{dt}\right|_{\text{adv}} = -\sum_{j\in\mathcal{N}(i)} q_{ij}. \tag{29}$$

Then,

$$\frac{d}{dt}\sum_{i=1}^{N} x_i = -\sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} q_{ij} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} (q_{ij} + q_{ji}) = 0. \tag{30}$$

Thus, in an idealized closed system with symmetric diffusion fluxes and antisymmetric advection fluxes, diffusion and advection redistribute pollutant mass across nodes without changing the total mass.

Real-world air-quality monitoring networks are not closed systems. Observed pollutant concentration is affected by external emissions, regional inflow, deposition, chemical loss, and unobserved boundary exchange. Moreover, the practical graph operators used in neural forecasting may involve row normalization, node-wise gates, and upwind directional discretization, which need not satisfy exact pairwise conservation. Therefore, the proposed physical anchor augments diffusion–advection dynamics with the learnable open-boundary source–sink term introduced in Eq. (5). The coefficient β captures net source–sink effects that cannot be represented by redistribution among observed stations alone. This formulation uses the closed-system conservation analysis as a motivating reference, while the implemented open-boundary anchor is not required to be exactly mass-conserving.

B.3 Topology Constraints of the Geospatial Graph

To accurately model the adjacency matrices ($\mathbf{A}^{\text{spa}}$ and $\mathbf{A}^{\text{adv}}$), we impose geographic constraints that restrict physical propagation across distant stations or high-elevation barriers. Following the spatial operator in the main text, the implemented topographic mask uses an absolute elevation threshold along the interpolated path between two stations.

Let ρ_i and ρ_j denote the geocoordinates of stations i and j, and let $d_{ij} = \|\rho_i - \rho_j\|$ be their geographic distance. We define the maximum path elevation as

$$e_{ij}^{\max} = \sup_{\lambda\in[0,1]} E\big(\lambda\rho_i + (1-\lambda)\rho_j\big), \tag{31}$$

where $E(\cdot)$ denotes the terrain elevation function. The base geospatial connectivity is then defined as

$$\mathbf{A}_{ij}^{\text{base}} = \mathbb{I}(d_{ij} < \theta_d) \cdot \mathbb{I}(e_{ij}^{\max} < \tau_h), \tag{32}$$

where θ_d is the distance threshold and τ_h is the absolute topographic elevation threshold, consistent with Eq. (8) in the main text.

This construction removes edges between stations that are either too far apart or whose interpolated path crosses terrain above the specified elevation threshold. The resulting mask is used to constrain spatial message passing so that pollutant diffusion is not propagated across implausible geophysical barriers.

C Extended Review of Related Work

In this section, we provide a more comprehensive discussion on the two foundational paradigms of air quality forecasting: the physics-based numerical branch and the deep learning-based data-driven branch.

C.1 Physics-Based Branch

These approaches rely on domain knowledge to construct Ordinary and Partial Differential Equations (ODEs/PDEs) that describe the physical mechanisms of air pollutant transport, such as diffusion and advection (Li et al., 2023; Sokhi et al., 2021). By solving these equations via numerical techniques like finite difference or finite element methods (Özişik et al., 2017; Ŝolín, 2005), these models theoretically capture the dispersion dynamics. However, they frequently face two major limitations:

1. They often rely on idealized closed-system assumptions that fail to capture real-world open-boundary dynamics;
2. The reliance on heavy numerical simulations becomes computationally prohibitive at large spatial scales, hindering rapid decision-making for urban policy (Lai et al., 2023).

C.2 Deep Learning Branch

Conversely, data-driven approaches have gained prominence by leveraging historical observations and auxiliary covariates (e.g., meteorology, POI distributions) to extract high-dimensional spatiotemporal representations (Han et al., 2021, 2023). While early shallow learning methods (Sánchez et al., 2013; Su et al., 2023) showed limited capacity, deep learning architectures—including RNNs, Transformers, and GNNs—have significantly enhanced performance by modeling complex non-linear dependencies (Yu et al., 2018; Luo et al., 2019; Liang et al., 2023; Zhou et al., 2024). Despite their power, these 'black-box' models often overlook essential physical laws, leading to confused spatiotemporal correlations that may violate fundamental physical principles, particularly under distribution shifts (Karpatne et al., 2017).

D Theoretical Stability Analysis of the Physical Anchor

The open-boundary term $\beta \odot \hat{X}^P$ in Eq. (5) introduces learnable source–sink effects into the physical anchor. Although this improves the expressiveness of the physical estimator, it also raises the question of whether the physical trajectory remains bounded over the forecasting horizon. We therefore provide a Lyapunov-type finite-horizon boundedness analysis. Unlike an idealized diffusion–advection analysis, the following derivation does not require the gated diffusion operator to be symmetric positive semi-definite or the upwind advection operator to be skew-symmetric.

D.1 Operator Form of the Physical Dynamics

Let $\mathbf{x}(t) \in \mathbb{R}^N$ denote the physical-anchor state $\hat{X}^P(t)$. For a fixed graph and meteorological condition, the physical dynamics can be written in the following linear operator form:

$$\frac{d\mathbf{x}}{dt} = \mathcal{H}(t)\mathbf{x}, \qquad \mathcal{H}(t) = \mathcal{H}_{\mathrm{diff}}(t) + \mathcal{H}_{\mathrm{adv}}(t) + \mathbf{B}, \tag{33}$$

where $\mathcal{H}_{\mathrm{diff}}(t)$ denotes the gated diffusion operator, $\mathcal{H}_{\mathrm{adv}}(t)$ denotes the gated upwind advection operator, and $\mathbf{B} = \mathrm{diag}(\beta)$ is the open-boundary source–sink operator.

Since the monitoring graph is irregular and the advection operator is direction-dependent, $\mathcal{H}_{\mathrm{diff}}(t)$ and $\mathcal{H}_{\mathrm{adv}}(t)$ are not necessarily symmetric. We therefore analyze the energy growth through the symmetric part of the full operator:

$$\mathrm{Sym}\big(\mathcal{H}(t)\big) = \frac{1}{2}\left(\mathcal{H}(t) + \mathcal{H}(t)^\top\right). \tag{34}$$

The corresponding Euclidean logarithmic norm is

$$\mu_2\big(\mathcal{H}(t)\big) = \lambda_{\max}\left(\mathrm{Sym}\big(\mathcal{H}(t)\big)\right), \tag{35}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue.

D.2 Lyapunov-Type Finite-Horizon Boundedness

We define the quadratic energy function

$$V(t) = \frac{1}{2}\|\mathbf{x}(t)\|_2^2. \tag{36}$$

Theorem D.1 (Finite-Horizon Boundedness of Open-Boundary Dynamics)**.** *Consider the physical-anchor dynamics in Eq.* (33)*. Assume that the induced operator has a bounded logarithmic norm over the forecasting horizon:*

$$\mu_2\big(\mathcal{H}(t)\big) \le \Gamma, \qquad \forall t \in [0, \tau], \tag{37}$$

for some finite constant Γ*. Then the physical trajectory satisfies*

$$\|\mathbf{x}(t)\|_2 \le \|\mathbf{x}(0)\|_2 \exp(\Gamma t), \qquad \forall t \in [0, \tau]. \tag{38}$$

Thus, the physical anchor does not exhibit finite-time blow-up over the prediction horizon.

Proof. Differentiating the Lyapunov energy gives

$$\begin{aligned} \dot{V}(t) &= \frac{d}{dt}\left(\frac{1}{2}\mathbf{x}(t)^\top \mathbf{x}(t)\right) \\ &= \mathbf{x}(t)^\top \frac{d\mathbf{x}(t)}{dt} \\ &= \mathbf{x}(t)^\top \mathcal{H}(t)\mathbf{x}(t). \end{aligned} \tag{39}$$

For any real matrix $\mathcal{H}(t)$, the quadratic form depends only on its symmetric part:

$$\mathbf{x}(t)^\top \mathcal{H}(t)\mathbf{x}(t) = \mathbf{x}(t)^\top \mathrm{Sym}\big(\mathcal{H}(t)\big)\mathbf{x}(t). \tag{40}$$

By the Rayleigh quotient,

$$\mathbf{x}(t)^\top \mathrm{Sym}\big(\mathcal{H}(t)\big)\mathbf{x}(t) \le \lambda_{\max}\left(\mathrm{Sym}\big(\mathcal{H}(t)\big)\right)\|\mathbf{x}(t)\|_2^2. \tag{41}$$

Using the definition of $\mu_2(\cdot)$ and the assumption $\mu_2(\mathcal{H}(t)) \le \Gamma$, we obtain

$$\dot{V}(t) \le \Gamma\|\mathbf{x}(t)\|_2^2 = 2\Gamma V(t). \tag{42}$$

Applying the Gronwall-Bellman inequality yields

$$V(t) \le V(0)\exp(2\Gamma t). \tag{43}$$

Taking the square root gives

$$\|\mathbf{x}(t)\|_2 \le \|\mathbf{x}(0)\|_2 \exp(\Gamma t). \tag{44}$$

This completes the proof. □

Physical Interpretation. This result establishes finite-horizon boundedness of the physical anchor under bounded operator growth. It does not assume that the gated diffusion operator is symmetric positive semi-definite, nor that the upwind advection operator is skew-symmetric. Instead, the energy growth is controlled by the largest eigenvalue of the symmetric part of the full diffusion–advection–source operator. In practice, bounded transport coefficients, regularization on the source–sink term β, numerical ODE integration over a finite horizon, and validation-based training jointly help keep the physical trajectory well behaved.

E Mathematical Formulation and Analysis of Spectral Reasoning

In this section, we provide the rigorous mathematical construction of the Spectral Operator used in the S^2 Reasoner, followed by a theoretical analysis of its global receptive field and filtering properties.

E.1 Spectral Micro-Operator Construction

Unlike spatial operators that aggregate local neighbors (e.g., GAT), the Spectral Operator manipulates signals in the orthogonal frequency domain. Let $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$ denote the spatial feature tensor, where H, W are the grid dimensions and C is the channel depth.

2D Discrete Fourier Transform (DFT). We first map the spatial signal to the frequency domain using the 2D DFT. For each channel c, the spectral coefficient at frequency mode (u, v) is computed as:

$$\mathcal{F}(\mathbf{U})_{u,v,c} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{U}_{h,w,c} \cdot \exp\left(-2\pi i \left(\frac{hu}{H} + \frac{wv}{W}\right)\right), \tag{45}$$

where $i = \sqrt{-1}$ is the imaginary unit. This step decomposes the spatial field into orthogonal sinusoidal basis functions at different spatial frequencies.

Localized Perturbation and Spectral Support. Consider an idealized perturbation at a single grid location (h_0, w_0): $\Delta\mathbf{U}_{h,w,c} = a\delta_{h,h_0}\delta_{w,w_0}$, where δ denotes the Kronecker delta. By Eq. (45), its Fourier transform is

$$\begin{aligned} \mathcal{F}(\Delta\mathbf{U})_{u,v,c} &= a \exp\left(-2\pi i \left(\frac{h_0 u}{H} + \frac{w_0 v}{W}\right)\right), \\ |\mathcal{F}(\Delta\mathbf{U})_{u,v,c}| &= |a|. \end{aligned} \tag{46}$$

Thus, a localized impulse contributes to all spatial frequency modes, rather than only high-frequency components. Although real pollution events have more complex spectra, this observation motivates modeling interactions across frequency modes.

Frequency-Domain Transformer Processing. After the FFT, the spectral coefficients are treated as a sequence of frequency tokens. We concatenate the real and imaginary parts of each complex coefficient and feed the resulting tokens into a lightweight Transformer encoder:

$$\widehat{\mathbf{S}} = \mathrm{Concat}\big(\mathrm{Re}(\widehat{\mathbf{U}}), \mathrm{Im}(\widehat{\mathbf{U}})\big), \tag{47}$$

$$\begin{aligned} \widehat{\mathbf{S}}' &= \mathrm{LayerNorm}\left(\widehat{\mathbf{S}} + \mathrm{MSA}(\widehat{\mathbf{S}})\right), \\ \widehat{\mathbf{S}}'' &= \mathrm{LayerNorm}\left(\widehat{\mathbf{S}}' + \mathrm{FFN}(\widehat{\mathbf{S}}')\right). \end{aligned} \tag{48}$$

The processed tokens $\widehat{\mathbf{S}}''$ are then reshaped back to the spectral grid before inverse FFT. This nonlinear frequency-domain processing allows the model to learn adaptive interactions among frequency modes, rather than relying on a fixed pointwise complex-valued filter.

Inverse DFT (Spatial Recovery). The modulated spectral features are transformed back to the spatial domain via the Inverse DFT:

$$\mathbf{U}'_{h,w,c} = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \widehat{\mathbf{U}}'_{u,v,c} \cdot \exp\left(2\pi i \left(\frac{hu}{H} + \frac{wv}{W}\right)\right). \tag{49}$$

By the duality of Fourier analysis, this operation distributes the global information captured in the frequency domain back to every local coordinate (h, w), achieving instantaneous long-range interaction.

E.2 Theoretical Analysis: Global Mixing View

We provide a linear spectral-filter view to clarify why frequency-domain processing naturally induces global spatial mixing. This analysis characterizes the FFT-based mixing backbone used in the spectral pathway. The full Frequency-Domain Transformer used in PAS^2 further introduces nonlinear self-attention and feed-forward transformations, and therefore should not be interpreted as being exactly equivalent to a linear convolutional filter.

Theorem E.1 (Global Mixing of Linear Spectral Filtering)**.** *Consider the linear spectral filtering subcase where grid features $\mathbf{U} \in \mathbb{R}^{H \times W \times d}$ are transformed into the frequency domain and modulated by a learnable frequency response $\mathcal{K}_\theta$:*

$$\widehat{\mathbf{U}} = \mathcal{F}_{\mathrm{FFT}}(\mathbf{U}), \qquad \widehat{\mathbf{U}}' = \mathcal{K}_\theta \odot \widehat{\mathbf{U}}, \qquad \mathbf{U}' = \mathcal{F}^{-1}_{\mathrm{FFT}}(\widehat{\mathbf{U}}'). \tag{50}$$

Then, for each channel, this operation is equivalent to a circular convolution in the spatial domain. Specifically, there exists a spatial kernel $\mathbf{G}_\theta = \mathcal{F}_{\mathrm{FFT}}^{-1}(\mathcal{K}_\theta)$ *such that*

$$\mathbf{U}'_{h,w} = (\mathbf{U} *_{\mathrm{circ}} \mathbf{G}_\theta)_{h,w} = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} \mathbf{U}_{m,n} \cdot \mathbf{G}_{\theta,(h-m) \bmod H,(w-n) \bmod W}. \tag{51}$$

Proof. By the convolution theorem, pointwise multiplication in the frequency domain corresponds to circular convolution in the spatial domain:

$$\mathcal{F}_{\mathrm{FFT}}\left(\mathbf{U} *_{\mathrm{circ}} \mathbf{G}_\theta\right) = \mathcal{F}_{\mathrm{FFT}}(\mathbf{U}) \odot \mathcal{F}_{\mathrm{FFT}}(\mathbf{G}_\theta). \tag{52}$$

Letting $\mathcal{K}_\theta = \mathcal{F}_{\mathrm{FFT}}(\mathbf{G}_\theta)$ gives $\widehat{\mathbf{U}}' = \mathcal{K}_\theta \odot \widehat{\mathbf{U}}$. Applying $\mathcal{F}_{\mathrm{FFT}}^{-1}$ yields the stated circular convolution form. □

Implication. The theorem shows that even the linear spectral-filter subcase has a global spatial mixing effect, since the induced circular convolution kernel can cover the entire $H \times W$ grid in a single spectral operation. In the implemented S^2 Reasoner, the Frequency-Domain Transformer generalizes this view by learning nonlinear interactions among frequency tokens rather than using only a fixed pointwise spectral response.

E.3 Resolution of the Low-Pass Dilemma

Standard spatial aggregators (e.g., Laplacian smoothing) act as low-pass filters. Let λ be the eigenvalues of the graph Laplacian. GCN propagation dampens high frequencies as $(1-\theta\lambda)^k \to 0$ for large λ. In contrast, our Spectral Operator learns the filter $\mathcal{K}_{u,v}$ directly. The network is unconstrained and can assign large weights $|\mathcal{K}_{u,v}| \gg 1$ to high-frequency modes $(u,v)_{\text{high}}$ that correspond to rapid spatial fluctuations ("Sudden Changes"). This provides a mechanism for preserving high-frequency residual variations, which is empirically supported by the ablation results on sudden-change subsets.

F Dataset Profiling and Experimental Protocols

In this section, we provide dataset statistics, pre-processing protocols, and visual analyses of the two primary benchmarks: the city-level **Beijing** dataset and the national-level **KnowAir** dataset.

F.1 Study Areas and Monitoring Networks

Figure 6 visualizes the spatial coverage of the two evaluation benchmarks. KnowAir provides a national-scale monitoring network spanning 184 major cities across mainland China, while the Beijing dataset focuses on a dense urban-scale monitoring network. The two settings therefore evaluate PAS^2 under complementary spatial regimes: long-range regional transport and fine-grained urban pollution dynamics.

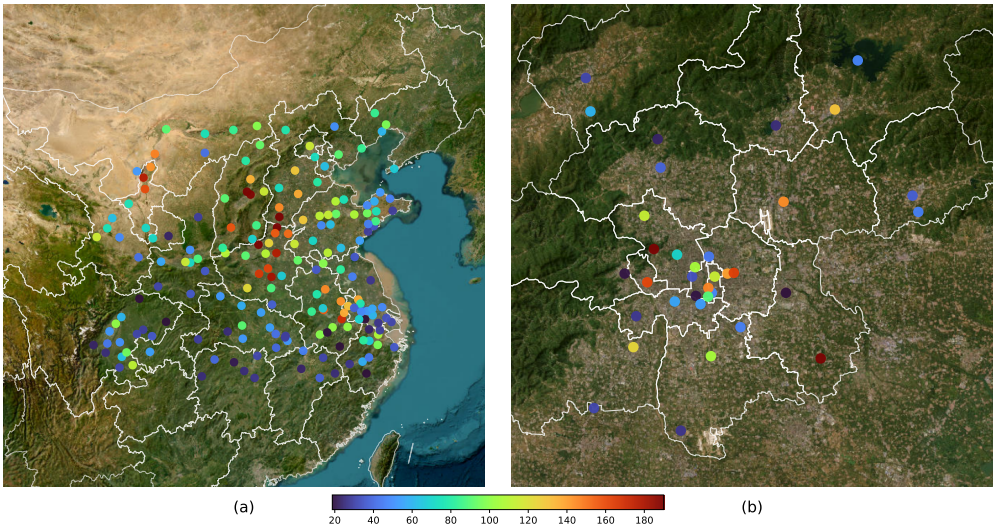


Figure 6: Geospatial distribution of the air quality monitoring networks. (a) The national-level KnowAir dataset covers 184 major cities across mainland China. (b) The Beijing dataset represents a dense urban-scale monitoring network.

F.2 Experimental Setup and Definitions

Forecasting Protocol. Following standard settings in prior work (Hettige et al., 2024; Liang et al., 2023), we aggregate raw hourly observations into a 3-hour resolution to reduce sensor-level noise. We use a historical window of $T = 24$ steps, corresponding to 72 hours, to predict the future trajectory of $\tau = 24$ steps. Future meteorological inputs $\hat{A}_{T+1:T+\tau}$ are predicted from $A_{1:T}$ only. No ground-truth meteorological observations from the forecast horizon are used, and all meteorology-based baselines receive the same predicted inputs. Missing values are imputed using the preceding 24-hour mean to preserve diurnal periodicity.

Definition of Sudden Change Scenarios. A core goal of PAS^2 is to capture high-volatility pollution events. Following Liang et al. (2023), we define a *Sudden Change* scenario as a time step satisfying two conditions:

1. **High Intensity**: absolute $\mathrm{PM}_{2.5} > 75\,\mu\mathrm{g/m}^3$.
2. **High Volatility**: $|\Delta\mathrm{PM}_{2.5}| > 20\,\mu\mathrm{g/m}^3$ within one 3-hour interval.

This subset isolates non-stationary high-gradient events and serves as a targeted testbed for evaluating the S^2 Reasoner.

An event refers to a station–time transition satisfying both criteria. The event ratio is the proportion of such transitions, while the window ratio is the proportion of forecasting windows containing at least one event. Table 6 reports the total event counts and the train/test ratios.

Table 6: Sudden-change statistics. Train/test entries show event ratio / window ratio (%).

Dataset	Events	Train (%)	Test (%)
Beijing	2,324	10.28 / 93.53	10.70 / 94.54
KnowAir	31,024	9.24 / 98.65	10.45 / 98.73

F.3 Beijing Dataset: Urban Micro-Scale Analysis

The Beijing dataset contains a fine-grained urban monitoring network with 35 stations from January 2017 to May 2018. It represents a dense urban scenario with strong anthropogenic influence and rapid local pollution variation.

F.3.1 Feature Statistics and Heavy-Tail Distribution

We analyze the feature distributions in Figure 7. The $PM_{2.5}$ distribution is right-skewed with a heavy tail, indicating frequent extreme smog episodes. The meteorological covariates also exhibit heterogeneous and skewed distributions, motivating robust normalization and regime-aware modeling.

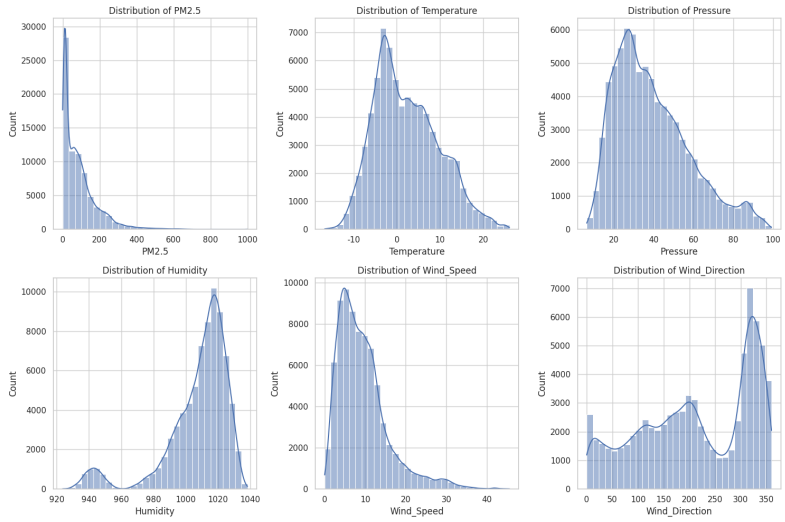


Figure 7: Feature distributions in Beijing. $PM_{2.5}$ exhibits a heavy-tailed distribution, while meteorological covariates show heterogeneous marginal statistics.

F.3.2 Urban Multivariate Correlations

Figure 8 presents the correlation structure between pollutants and meteorological variables. The negative correlation between wind speed and $PM_{2.5}$ is consistent with the physical intuition that stagnant air facilitates accumulation, whereas stronger ventilation promotes dispersion.

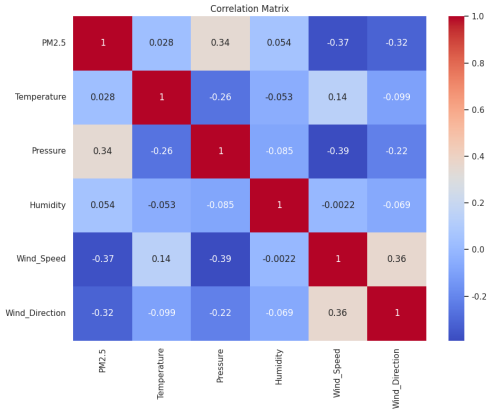


Figure 8: Pearson correlation matrix for the Beijing dataset, highlighting the dependency between ventilation-related variables and pollutant accumulation.

F.4 KnowAir Dataset: Regional Macro-Scale Analysis

The KnowAir dataset covers 184 cities across mainland China from January 2015 to December 2018. It provides a national-scale benchmark with broader spatial coverage and richer reanalysis meteorology, testing the model's capacity to capture long-range transport dependencies.

F.4.1 Long-Term Seasonal Trends

Figure 9 depicts the long-term evolution of mean $PM_{2.5}$ concentration in KnowAir from 2015 to 2018. The series shows clear seasonal variation, with recurrent winter peaks and lower summer concentrations. The intermittent high-magnitude spikes further indicate abrupt pollution episodes, motivating models that capture both persistent temporal patterns and short-term volatility.

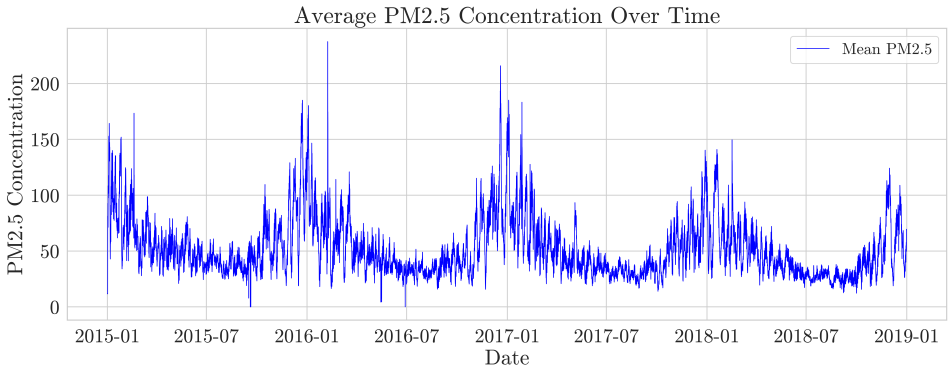


Figure 9: Long-term evolution of mean $PM_{2.5}$ concentration in KnowAir from 2015 to 2018.

F.4.2 Advanced Meteorological Dependencies

Unlike Beijing, KnowAir includes Boundary Layer Height (BLH), which characterizes vertical mixing conditions. As shown in Figure 10, BLH is negatively correlated with $PM_{2.5}$, consistent with the physical intuition that a lower boundary layer restricts vertical dispersion and concentrates near-surface aerosols.

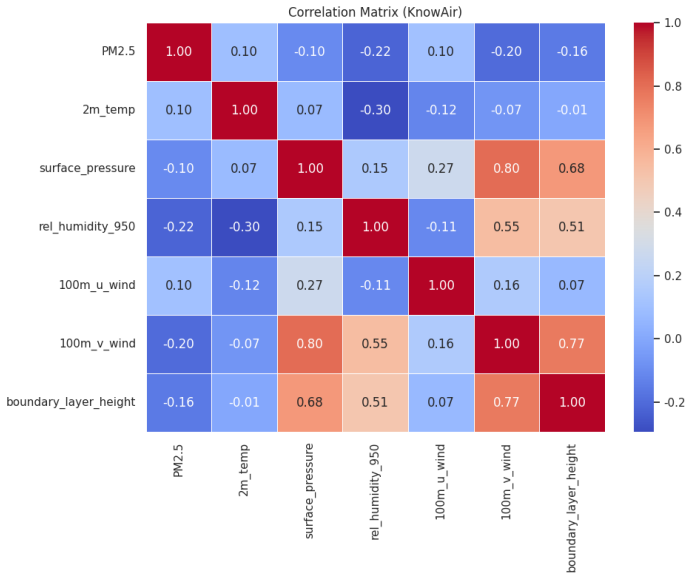


Figure 10: Correlation matrix for KnowAir. The negative correlation between BLH and $PM_{2.5}$ highlights the role of vertical atmospheric structure in pollutant accumulation.

F.4.3 Spatial Non-Stationarity

Figure 11 displays the station-level variability of $PM_{2.5}$ in KnowAir. The differences in medians, dispersion ranges, and outlier frequencies demonstrate strong spatial non-stationarity across cities. This heterogeneity supports the use of station-specific physical context and regime-adaptive routing rather than globally shared dynamics.

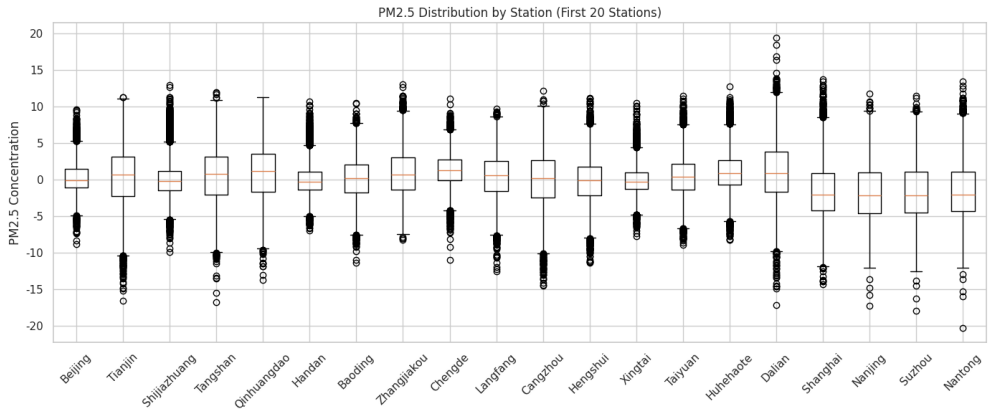
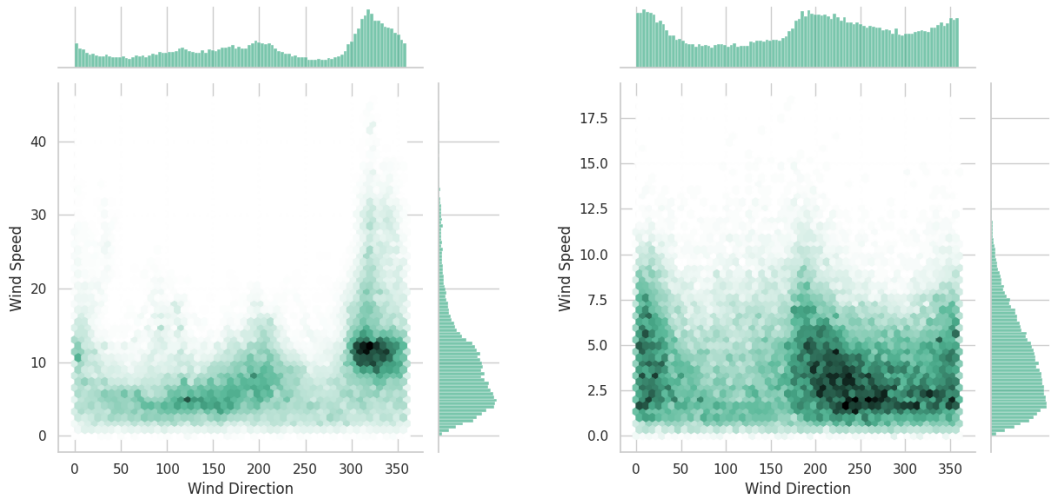


Figure 11: Station-level $PM_{2.5}$ variability in KnowAir. The heterogeneous distributions motivate spatially aware modeling of local transport regimes.

F.5 Meteorological Forcing and Wind Regimes

Wind speed and wind direction are key meteorological drivers of advective pollutant transport. To characterize the wind climatology of the two study areas, Figure 12 visualizes the joint empirical distributions of wind direction and wind speed for Beijing and KnowAir. Darker hexagonal bins indicate more frequent wind direction–speed regimes, while the marginal histograms summarize the corresponding one-dimensional distributions.



(a) Beijing dataset

(b) KnowAir dataset

Figure 12: Joint empirical distributions of wind direction and wind speed in the Beijing and KnowAir datasets. Darker hexagonal bins correspond to more frequent wind direction–speed regimes.

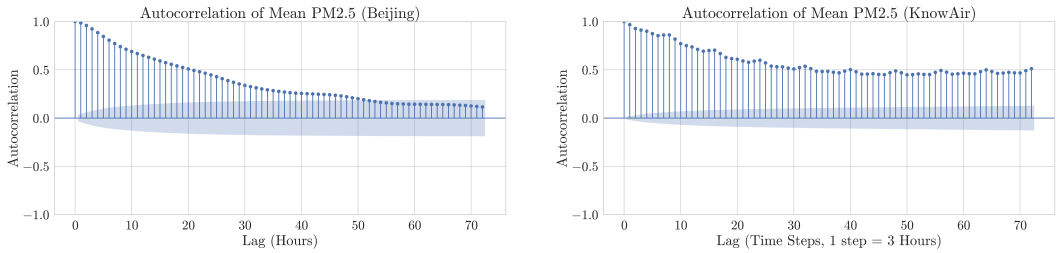
Figure 12 shows that both datasets exhibit non-uniform wind regimes rather than isotropic transport conditions. Beijing is dominated by low-to-moderate wind-speed regimes across several directions, while KnowAir exhibits broader directional variability and stronger high-speed directional clusters. These observations support the use of a direction-aware advection operator in the physical anchor, allowing the model to account for anisotropic pollutant transport under heterogeneous meteorological forcing.

G Extended Meteorological Analysis

In addition to basic feature distributions, we perform advanced signal processing and meteorological analysis to characterize the datasets.

G.1 Temporal Autocorrelation Analysis

Air quality sequences exhibit temporal persistence and periodicity. Figure 13 reports the autocorrelation function (ACF) of mean $PM_{2.5}$ concentrations for Beijing and KnowAir. Since all observations are aggregated into 3-hour intervals, a lag of 8 corresponds to one day.



(a) Beijing dataset (b) KnowAir dataset

Figure 13: Autocorrelation of mean $PM_{2.5}$ concentrations. The x-axis denotes lag in 3-hour time steps. Both datasets exhibit strong short-term persistence, while KnowAir shows more pronounced long-range periodicity.

The ACF curves indicate that $PM_{2.5}$ dynamics contain both short-term temporal persistence and longer-range periodic structure. This supports the use of continuous-time dynamics for modeling smooth temporal evolution, while also motivating additional mechanisms for abrupt deviations from these persistent trends.

G.2 Wind Rose Analysis

Wind direction is a key meteorological driver of advective pollutant transport. Figure 14 visualizes the empirical wind-direction distributions of the Beijing and KnowAir datasets using polar frequency histograms. The observed directional biases indicate that pollutant transport is strongly anisotropic, motivating the use of a direction-aware advection operator in the physical anchor.



(a) Beijing dataset (b) KnowAir dataset

Figure 14: Polar frequency histograms of wind direction in the Beijing and KnowAir datasets. The non-uniform directional distributions reflect anisotropic transport regimes, supporting the design of the advection component in the physical anchor.

G.3 Station Heterogeneity Analysis

To further demonstrate the complexity of the graph network, we analyze the variance of PM2.5 distributions across different monitoring stations (Figure 15).

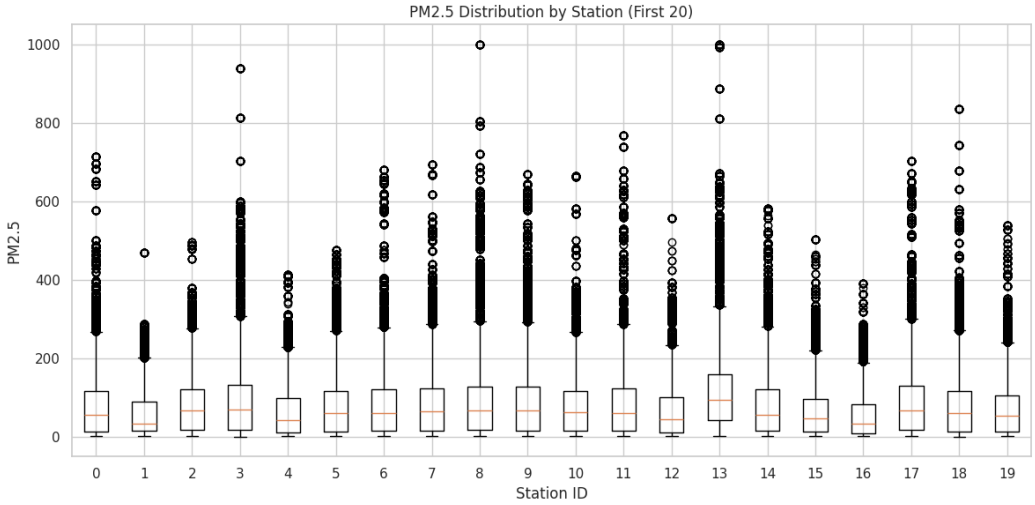


Figure 15: Boxplots of PM2.5 concentrations for the top 20 stations in the Beijing dataset. The variability in interquartile ranges (box height) and outliers (dots) highlights the spatial non-stationarity of the pollution field.

The significant differences in median values and spread between stations suggest that a global model parameters might be insufficient. This observation supports the design choice of PAS2 to incorporate Node-Adaptive Parameters (NAP) or similar graph mechanisms to learn station-specific embeddings.

H Evaluation Metrics

To comprehensively assess the predictive performance of the proposed PAS^2 framework, we employ three widely-adopted regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE).

Let $X_{T+t,i}$ denote the ground-truth pollutant concentration for the i-th monitoring station at the future time step $T+t$, and $\hat{X}_{T+t,i}$ denote the corresponding predicted value. The evaluation is conducted across all N stations in the geospatial graph and the entire prediction horizon τ. The metrics are formulated as follows:

Mean Absolute Error (MAE) measures the average absolute deviation between the predictions and the ground truth, providing a direct interpretation of the prediction error magnitude:

$$\text{MAE} = \frac{1}{\tau N} \sum_{t=1}^{\tau} \sum_{i=1}^{N} \left| X_{T+t,i} - \hat{X}_{T+t,i} \right|. \tag{53}$$

Root Mean Square Error (RMSE) utilizes a quadratic scoring rule. By squaring the errors before averaging, RMSE assigns significantly larger penalties to major outliers. This metric is particularly crucial for our study, as it effectively evaluates the model's capability to capture extreme fluctuations and high-concentration peaks:

$$\text{RMSE} = \sqrt{\frac{1}{\tau N} \sum_{t=1}^{\tau} \sum_{i=1}^{N} \left(X_{T+t,i} - \hat{X}_{T+t,i} \right)^2}. \tag{54}$$

Symmetric Mean Absolute Percentage Error (SMAPE) measures the relative error percentage. Unlike standard MAPE, which becomes ill-posed when ground truth values are close to zero, SMAPE

bounds the error within $[0, 2]$, ensuring numerical stability across different background concentration regimes:

$$\text{SMAPE} = \frac{1}{\tau N} \sum_{t=1}^{\tau} \sum_{i=1}^{N} \frac{2 \left| X_{T+t,i} - \hat{X}_{T+t,i} \right|}{|X_{T+t,i}| + \left| \hat{X}_{T+t,i} \right|}. \tag{55}$$

I Experimental Configuration Details

To facilitate reproducibility, we provide the detailed experimental configurations utilized in this study. Table 7 summarizes the optimal hyperparameter settings for the PAS^2 framework. These values were rigorously selected via grid search on the validation sets of both the Beijing and KnowAir benchmarks. All experiments were conducted on a computational cluster equipped with four NVIDIA GeForce RTX 4090 GPUs. Each training run used a single RTX 4090 GPU with approximately 12 GB peak GPU memory. The average training time of PAS^2 was approximately 3.5 hours on KnowAir and 0.8 hours on Beijing. The full set of reported experiments required approximately 220 GPU-hours in total, excluding preliminary exploratory runs.

Table 7: Hyperparameter Configuration for the PAS^2 Framework.

CATEGORY	HYPERPARAMETER	VALUE
Data Processing	Input Sequence Length (T_{in})	24 (72h)
	Prediction Horizon (T_{out})	24 (72h)
	Input Dimensions	6 (Beijing) / 18 (KnowAir)
	Data Normalization	Z-score
	Split Ratio (Train/Val/Test)	7:1:2
Architecture	Node Embedding Dim (d_{node})	10
	Time Embedding Dim (d_{time})	32
	Encoder Hidden Units	64
	ODE Solver Hidden Units	64
	Graph Layers	2
	Activation Function	Tanh / ReLU
Optimization	Optimizer	Adam
	Learning Rate (η)	1×10^{-3}
	Scheduler	StepLR
	Decay Gamma	0.5 (Step=50)
	Batch Size	64
	Early Stopping Patience	20 epochs
ODE Solver	Method	RK4 / Euler
	Integration Step Size	0.1
	Relative Tolerance	1×10^{-5}

J Additional Performance Analysis

J.1 Statistical Stability under Randomized Trials

To evaluate the statistical stability of the main results, we repeat the experiments over five independent runs with different random seeds. The randomness comes from mini-batch sampling and parameter initialization, while the train/validation/test splits are kept fixed. Table 8 reports the mean and standard deviation for PAS^2 and the top-3 strongest baselines on each metric.

Table 8: Performance comparison under randomized trials. We report mean ± standard deviation over 5 independent runs. For each metric, PAS^2 and the top-3 strongest baselines are reported. The best results are highlighted in red.

Dataset	Setting	Metric	PAS^2	Best Baseline	2nd / 3rd Baseline
KnowAir	3days	MAE	18.70 ± 0.05	18.89 ± 0.06 (AirFormer)	19.06 ± 0.08 (Air-DualODE) / 19.21 ± 0.11 (ASTGCN)
		RMSE	29.53 ± 0.07	29.79 ± 0.12 (Air-DualODE)	29.83 ± 0.12 (AirFormer) / 30.73 ± 0.15 (ASTGCN)
		SMAPE	0.420 ± 0.004	0.430 ± 0.004 (ASTGCN)	0.430 ± 0.004 (Air-DualODE) / 0.440 ± 0.005 (AirFormer)
	Sudden Change	MAE	48.41 ± 0.28	50.59 ± 0.36 (ASTGCN)	51.15 ± 0.47 (AirFormer) / 51.32 ± 0.29 (Air-DualODE)
		RMSE	68.01 ± 0.35	71.26 ± 0.27 (AirFormer)	71.54 ± 0.52 (ODE-RNN) / 72.48 ± 0.45 (LatentODE)
		SMAPE	0.480 ± 0.005	0.510 ± 0.005 (Air-DualODE)	0.520 ± 0.008 (ODE-RNN) / 0.520 ± 0.006 (AirFormer)
Beijing	3days	MAE	31.24 ± 0.16	32.13 ± 0.14 (AirPhyNet)	33.97 ± 0.21 (AirFormer) / 36.81 ± 0.52 (STGCN)
		RMSE	68.29 ± 0.26	70.23 ± 0.39 (AirFormer)	71.81 ± 0.30 (AirPhyNet) / 73.35 ± 0.57 (STGCN)
		SMAPE	0.480 ± 0.005	0.530 ± 0.007 (AirPhyNet)	0.540 ± 0.011 (AirFormer) / 0.570 ± 0.016 (LatentODE)
	Sudden Change	MAE	148.46 ± 0.92	152.14 ± 1.02 (AirPhyNet)	153.04 ± 1.15 (AirFormer) / 156.50 ± 0.97 (Air-DualODE)
		RMSE	217.80 ± 1.18	220.72 ± 2.56 (AirPhyNet)	221.70 ± 2.05 (AirFormer) / 222.35 ± 2.03 (STTN)
		SMAPE	0.720 ± 0.004	0.730 ± 0.006 (AirFormer)	0.740 ± 0.004 (AirPhyNet) / 0.770 ± 0.010 (STGCN)

J.2 Detailed Discussion of Overall Performance

We provide a more detailed discussion of the results in Table 1. First, deep learning approaches consistently outperform classical statistical methods such as HA and VAR, demonstrating their stronger capacity to model nonlinear spatiotemporal correlations in air pollution dynamics. Second, generic spatio-temporal models transfer reasonably well to air quality forecasting, but they generally lag behind domain-specific predictors such as AirFormer, suggesting that atmospheric inductive biases remain important for this task.

Third, pure Neural ODE-based approaches are competitive but generally inferior to strong spatio-temporal architectures. This indicates that continuous-time modeling alone is insufficient without explicit spatial dependency modeling across monitoring stations. Finally, among physics-guided approaches, PAS^2 consistently outperforms AirPhyNet and Air-DualODE. This supports the proposed physics-anchored residual formulation, where the physical estimator provides a macroscopic anchor and the neural branch focuses on residual correction, rather than treating the two branches as peer predictors.

The advantage of PAS^2 is particularly evident under *Sudden Change* scenarios. These cases contain abrupt high-concentration fluctuations that are easily oversmoothed by purely spatial aggregators. The improvement in this subset aligns with our design motivation: the S^2 Reasoner recovers high-frequency residual components through the spectral path, while RACE adaptively selects regime-specific dynamics under rapidly changing atmospheric conditions. Overall, the results suggest that physics anchoring and spatial-spectral residual reasoning jointly improve robustness in volatile pollution regimes.

K Latent Representation and Spectral Analysis

K.1 Latent Manifold Visualization

To interpret the internal dynamics of PAS^2, we extracted the high-dimensional latent states $z(t)$ from the ODE solver and projected them onto a 2D plane using t-SNE (t-Distributed Stochastic Neighbor Embedding). As shown in Figure 16, the model learns to effectively organize the dynamics into distinct regimes corresponding to different air quality levels (Clean, Moderate, Polluted) without explicit supervision labels. This clustering indicates that the continuous-time ODE formulation successfully captures the underlying phase transitions in the pollution diffusion process.

K.2 Learned Spatial Topology Preservation

We investigated whether the learned station embeddings ($E_{station} \in \mathbb{R}^d$) preserve the physical geographical topology. Figure 17 presents a side-by-side comparison of the physical station map and the PCA projection of the learned embeddings. Remarkably, the topological structure (e.g., the clustering of urban center stations and the relative positions of suburban stations) is preserved in

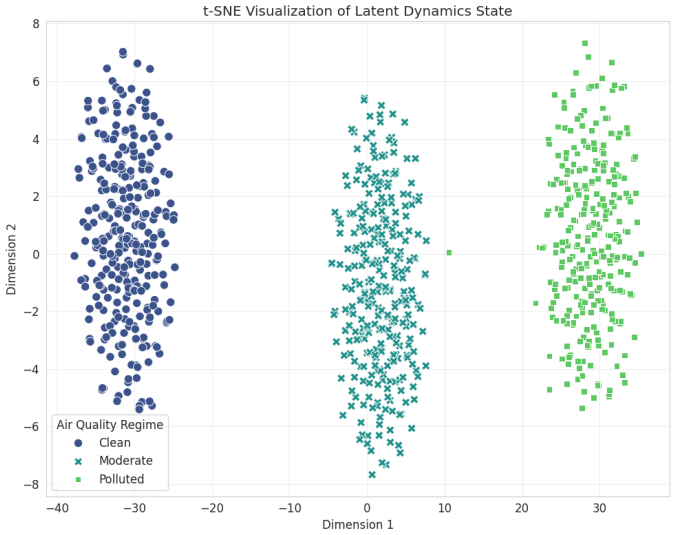


Figure 16: t-SNE visualization of the learned latent dynamics $z(t)$. The distinct clusters suggest the model has learned to disentangle different pollution regimes.

the learned latent space. This confirms that the AirEmbedding module encodes semantic spatial information that aligns with physical reality.

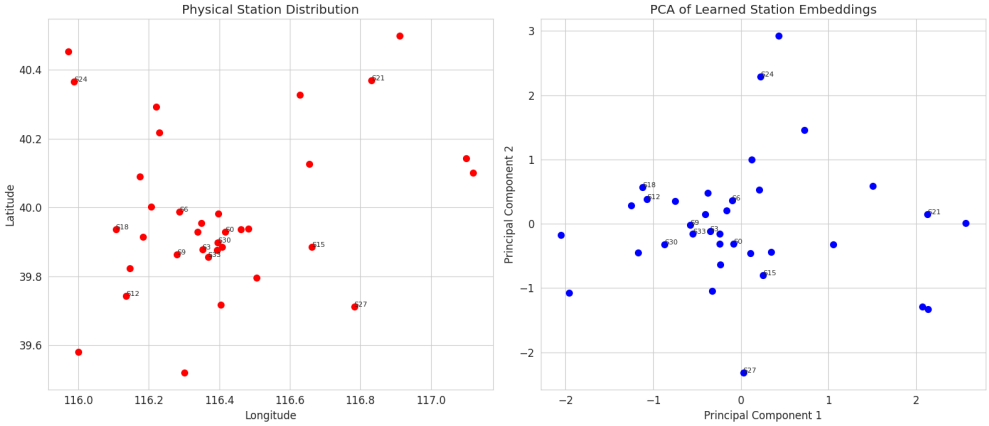


Figure 17: Topological Consistency Analysis: Physical Station Locations (Left) vs. PCA of Learned Embeddings (Right).

K.3 Spectral Feature Analysis

Air quality data exhibits strong multi-scale periodicities. To verify the efficacy of the Spectral ODE component, we analyzed the spectral properties of the modeled signals. Figure 18 displays the Power Spectral Density (PSD). The plot reveals prominent energy peaks at frequencies corresponding to daily ($T = 24h$) and weekly ($T = 168h$) cycles. The model's ability to reconstruct these high-frequency harmonics verifies that the spectral domain modeling prevents the "over-smoothing" problem common in standard Neural ODEs.

K.4 Neighbor Support of Sudden-Change Events

We examine sudden-change events at the original monitoring stations before grid interpolation. An event is considered neighbor-supported if at least one of its four nearest stations shows a concentration change in the same direction exceeding $10\,\mu\mathrm{g}/\mathrm{m}^3$ during the same three-hour interval.

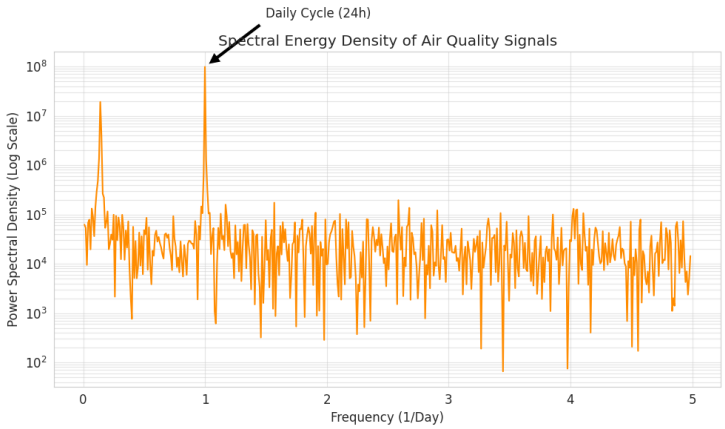


Figure 18: Power Spectral Density (PSD) analysis highlighting the capture of dominant temporal frequencies (Daily and Weekly cycles).

For comparison, we construct a shuffled control by sampling another timestamp from the same station and month, repeating the procedure 1,000 times. The observed neighbor-support rates are 84.9% on KnowAir and 78.4% on Beijing, compared with 41.1% and 42.5% under the shuffled control, respectively.

The higher observed rates indicate that many sudden changes are supported by nearby stations before interpolation. These results demonstrate that sudden-change events often exhibit spatially coherent pollutant variations across neighboring monitoring stations, even before grid interpolation.

K.5 Spectral Stability under Grid Projection

We evaluate the stability of spatial frequency patterns under different station-to-grid projections. Using the standard IDW projection as the reference, we compare the radial spectra obtained with RBF interpolation and nearest-neighbor interpolation. We also examine the effect of masking 20% of monitoring stations before projection. Table 9 reports the cosine similarity between each resulting spectrum and its reference.

Table 9: Cosine similarity of radial spectra under different projection settings, relative to the standard IDW projection.

Projection setting	KnowAir	Beijing
RBF interpolation	0.976	0.991
Nearest-neighbor interpolation	0.947	0.953
20% station masking	0.991	0.984

The high spectral similarities indicate that the spatial frequency patterns remain largely consistent across the tested interpolation methods and station-masking setting.

K.6 Spatiotemporal Attention Mechanism

Figure 19 visualizes the learned attention weights matrix. The heatmap reveals a sparse and non-uniform attention distribution. Specifically, we observe vertical bands indicating "Hub Nodes"—stations that exert a strong influence on many other stations (likely upwind sources). This interprets how the model aggregates spatial context for forecasting.

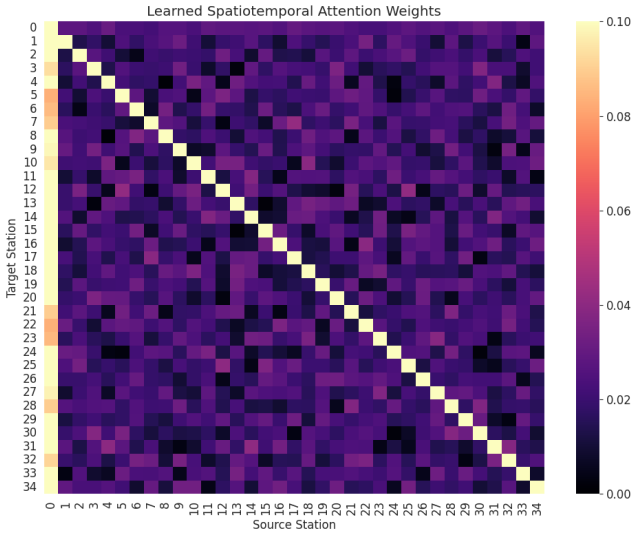


Figure 19: Learned Attention Weights Heatmap. Darker regions indicate stronger spatial dependencies between source and target stations.

L Training Algorithm

Algorithm 1 summarizes the training procedure of PAS^2. The key implementation detail is that the physical trajectory and physical context are detached before being passed to the neural residual branch, ensuring the optimization separation described in Section 3.4.

Algorithm 1 PAS^2 Training Procedure

Require: Batch historical concentrations $\mathbf{X}_b$, targets $\mathbf{Y}_b$, meteorology $\mathbf{A}_b$, graph G
Require: Parameters $\Theta = \{\Theta_{\text{phy}}, \Theta_{\text{net}}\}$
Ensure: Optimized parameters Θ^*
Initialize Θ_{phy} and Θ_{net}
Construct physical micro-operators L_{diff} and L_{adv}
for epoch $e = 1$ to E **do**
  for batch $(\mathbf{X}_b, \mathbf{Y}_b, \mathbf{A}_b)$ in DataLoader **do**
    // Physical anchoring
    $\hat{X}^P \leftarrow \text{ODESolve}(X_{b,T}, F_{\text{phy}}(\cdot; \Theta_{\text{phy}}, \mathbf{A}_b, G))$
    $C^P \leftarrow \text{ContextProj}(\hat{X}^P, \alpha, \beta)$
    $\bar{X}^P \leftarrow \text{sg}(\hat{X}^P), \quad \bar{C}^P \leftarrow \text{sg}(C^P)$
    // Neural residual reasoning
    $Z^R(0) \leftarrow \text{RACE_Encoder}(\mathbf{X}_b, \mathbf{A}_b, \bar{C}^P; \Theta_{\text{net}})$
    $Z^R(1) \leftarrow \text{ODESolve}(Z^R(0), \mathcal{F}_{S^2}(\cdot; \Theta_{\text{net}}, G, \bar{C}^P))$
    $\hat{X}^R \leftarrow \text{Decoder}(Z^R(1))$
    // Anchored prediction and separated losses
    $\hat{Y} \leftarrow \bar{X}^P + \hat{X}^R$
    $\mathcal{L}_{\text{pred}} \leftarrow \text{MSE}(\hat{Y}, \mathbf{Y}_b)$
    $\mathcal{L}_{\text{phy}} \leftarrow \text{MSE}(\hat{X}^P, \mathbf{Y}_b)$
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{pred}} + \lambda_1 \mathcal{L}_{\text{phy}} + \lambda_2 \mathcal{L}_{\text{reg}}$
    Update $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{\text{total}}$
  end for
end for
return Θ^*

M Mathematical Background: Neural ODEs

The core of our continuous-time modeling is based on the formulation of Neural Ordinary Differential Equations. Instead of specifying a discrete sequence of hidden layers $h_{t+1} = f(h_t, \theta)$, we parameterize the derivative of the hidden state using a neural network $f(h(t), t, \theta)$:

$$\frac{dh(t)}{dt} = f(h(t), t, \theta) \tag{56}$$

The hidden state at any future time t_1 can be computed from the initial state $h(t_0)$ by solving the initial value problem (IVP):

$$h(t_1) = h(t_0) + \int_{t_0}^{t_1} f(h(\tau), \tau, \theta) d\tau \tag{57}$$

In PAS2, we employ a numerical solver (e.g., Runge-Kutta 4th order, RK4) to approximate this integral. The RK4 update rule is given by:

$$k_1 = f(t_n, y_n) \tag{58}$$

$$k_2 = f(t_n + \frac{h}{2}, y_n + h\frac{k_1}{2}) \tag{59}$$

$$k_3 = f(t_n + \frac{h}{2}, y_n + h\frac{k_2}{2}) \tag{60}$$

$$k_4 = f(t_n + h, y_n + hk_3) \tag{61}$$

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{62}$$

where h is the integration step size. This continuous formulation allows us to handle irregular time samples and model the smooth evolution of atmospheric physics more naturally than discrete RNNs.

N Case Study Visualization

To visually assess the model's forecasting capability, we randomly selected a station from the Beijing dataset and visualized the prediction results over a 24-hour horizon.

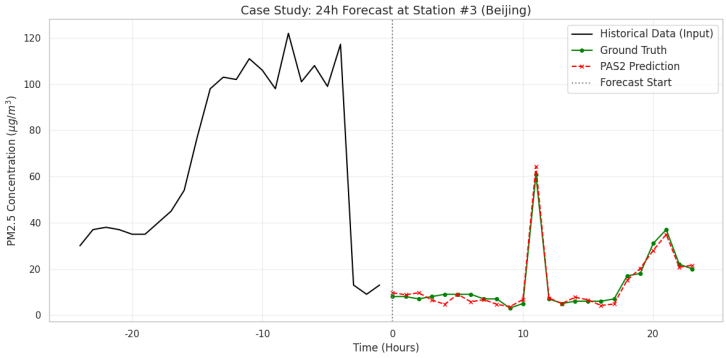


Figure 20: A case study visualization showing the 24-hour PM2.5 forecast at a representative station. The black line represents the historical input sequence (observed data), the green line shows the ground truth future values, and the red dashed line indicates the PAS2 model's prediction. The model successfully captures the rising trend and the subsequent stabilization, demonstrating its ability to model non-stationary pollutant dynamics.

As shown in Figure 20, the model tracks the "Ground Truth" trajectory closely, even when the trend shifts from the historical pattern. This validates the effectiveness of PAS^2 in learning the underlying differential dynamics of air pollution diffusion.

O Detailed Analysis of Baseline Models

To comprehensively evaluate the performance of PAS^2, we compared it against 14 representative baselines spanning four distinct categories: classical statistical methods, continuous-time ODE models, generic spatio-temporal deep learning models, and domain-specific physics-guided approaches. Below, we detail the implementation logic of these baselines and analyze their theoretical limitations in the context of high-volatility air quality forecasting.

O.1 Classical Statistical Methods

- **HA (Historical Average)**: Predicts future values based on the average of historical observations at the corresponding time steps. It serves as a static benchmark, failing to capture temporal trends.
- **VAR (Vector Auto-Regression)**: A generalized linear autoregressive model that captures pairwise relationships between multiple time series. While effective for short-term dependencies, it assumes linearity and cannot model complex non-linear atmospheric transport dynamics.

O.2 Continuous-Time Neural ODE Models

These models leverage differential equations to handle irregular time series but often lack explicit spatial modeling capabilities.

- **LatentODE** (Rubanova et al., 2019): Extends the Neural ODE (Chen et al., 2018) framework to process time series by encoding them into a latent trajectory. It excels at temporal continuity but treats spatial nodes independently, ignoring inter-station diffusion.
- **ODE-RNN** (Rubanova et al., 2019): A hybrid architecture combining Neural ODEs with GRU cells. It updates the hidden state continuously between observations but suffers from the same lack of explicit spatial reasoning as LatentODE.
- **ODE-LSTM** (Lechner & Hasani, 2020): Utilizes a Long Short-Term Memory (LSTM) network with a continuous-time hidden state evolution (LTC). It is designed for long-term dependencies but is computationally heavy and lacks graph topological awareness.

O.3 Generic Spatio-Temporal Graph Neural Networks

These data-driven models are the dominant paradigm in traffic and weather forecasting but often suffer from the "Low-Pass Dilemma."

- **DCRNN** (Li et al., 2018b): Models diffusion using a bidirectional random walk on the graph, integrated into an Encoder-Decoder architecture. It assumes a static diffusion process, which limits its adaptability to rapidly changing wind fields.
- **STGCN** (Yu et al., 2018): Combines spectral graph convolutions with 1D temporal convolutions. As noted in spectral graph theory, the GCN operator acts as a low-pass filter, inevitably smoothing out high-frequency "Sudden Change" signals critical for pollution peaks.
- **ASTGCN** (Guo et al., 2019): Enhances STGCN with spatial and temporal attention mechanisms. While attention improves interpretability, it still relies on local spatial aggregation that dilutes extreme values.
- **GTS** (Shang & Chen, 2021): Learns a discrete probabilistic graph structure simultaneously with forecasting. While it captures latent dependencies, the optimization of the discrete graph structure is unstable under high-volatility distributions.
- **GMAN** (Zheng et al., 2020): A graph multi-attention network that uses spatio-temporal attention blocks to capture correlations. Despite its strong performance, the $O(N^2)$ complexity of global attention limits its scalability and it struggles to model strict physical causality.
- **STTN** (Xu et al., 2020): A spatial-temporal transformer network that uses a sliding window attention mechanism. It focuses on long-range dependencies but lacks a mechanism to enforce physical consistency in mass transport.

Table 10: Comparison with modern time-series forecasting models on KnowAir. Lower values are better.

Model	Overall			Sudden Change		
	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
iTransformer	20.07	31.97	0.45	52.73	72.74	0.55
TimeMixer	20.45	32.45	0.46	52.22	71.98	0.54
PatchTST	20.52	32.67	0.46	53.98	73.91	0.57
TimesNet	20.84	32.82	0.47	52.95	72.80	0.55
FEDformer	23.09	35.01	0.54	53.55	73.18	0.56
PAS^2	**18.70**	**29.53**	**0.42**	**48.41**	**68.01**	**0.48**

O.4 Domain-Specific and Physics-Guided Approaches

These represent the current state-of-the-art (SOTA), incorporating domain knowledge or physical laws.

- **PM2.5-GNN** (Wang et al., 2020): A domain-knowledge-enhanced GNN that incorporates wind speed and direction as edge attributes. However, it still operates within a discrete graph framework and lacks a continuous physical evolution mechanism.
- **AirFormer** (Liang et al., 2023): A Transformer-based model specifically designed for air quality, utilizing Dartboard Attention to model multi-scale spatial dependencies. It is a pure data-driven model and lacks explicit physical constraints, leading to potential violation of conservation laws during sudden changes.
- **AirPhyNet** (Hettige et al., 2024): A physics-guided architecture that embeds fluid dynamics equations into the network. **Limitation:** It typically adopts a parallel coupling strategy and assumes a closed system (mass conservation), which contradicts the open-boundary nature of urban environments, limiting its accuracy in dissipative scenarios.
- **Air-DualODE** (Tian et al., 2025): The strongest baseline, employing dual Neural ODEs to model physical and residual terms. **Limitation:** As discussed in our methodology, it treats the physical and neural branches as peer predictors (Parallel Coupling), leading to gradient competition. Furthermore, it relies on spatial graph operators for the residual branch, failing to resolve the high-frequency transients that our proposed **Spectral Reasoner** effectively captures.

O.5 Modern Time-Series Forecasting Baselines

We additionally compare PAS^2 with five recent time-series forecasting models: PatchTST (Nie et al., 2022), TimesNet (Wu et al., 2022), FEDformer (Zhou et al., 2022), iTransformer (Liu et al., 2024), and TimeMixer (Wang et al., 2024). All models are evaluated on KnowAir under the same forecasting protocol. Table 10 reports the overall and sudden-change forecasting errors.

PAS^2 achieves the lowest errors across all six metrics in this comparison. The gains on sudden-change samples further support the value of physics-guided spatial-spectral reasoning for air-quality forecasting.

P Extended Ablation Analysis

Due to space constraints in the main text, we provide a more comprehensive discussion and mechanistic analysis of the ablation studies in this section.

The physics-only predictor uses $\hat{X}^P$ without the neural residual branch. Its 3-day and sudden-change MAEs are 24.88/61.71 on KnowAir and 45.36/181.32 on Beijing, compared with 18.70/48.41 and 31.24/148.46 for the full PAS^2, respectively. These results show the benefit of residual correction, particularly under sudden changes.

We also replace the physical ODE with a parameter-matched learned anchor, keeping the residual branch and training setup unchanged. The 3-day/sudden-change MAEs increase to 20.35/53.20 on

KnowAir and 35.10/161.80 on Beijing. This comparison supports the contribution of the prescribed transport operators beyond anchor capacity.

To examine the role of each physical operator, we remove diffusion, advection, or the source–sink term separately. We compare the relative increase in physics-only MAE between the bottom and top quartiles of the corresponding driver: spatial concentration gradient, wind-aligned transport, and non-conservative concentration change, respectively.

Table 11: Relative increase in physics-only MAE (%) after removing each operator. Each entry reports the low/high driver quartiles.

Removed operator	KnowAir	Beijing
Diffusion	3.2 / 12.1	3.7 / 10.8
Advection	2.9 / 14.6	3.4 / 13.1
Source–sink	3.0 / 10.7	3.2 / 9.6

Removing each operator causes a larger error increase when its corresponding driver is strong. This supports the distinct roles of the prescribed physical operators.

We further examine how the physical and neural branches behave under sudden changes. The neural-to-physical change ratio increases from 0.36 to 0.64 on KnowAir and from 0.40 to 0.68 on Beijing, comparing regular and sudden-change transitions. The residual-to-physical Dirichlet-energy ratio also increases from 2.2 to 2.9 on KnowAir and from 2.0 to 2.6 on Beijing, comparing overall and sudden-change samples. These results suggest stronger and less-smooth neural corrections during sudden changes.

P.1 Why Parallel Coupling Fails: Functional Redundancy and Gradient Competition

Most existing Physics-Guided Learning (PGL) methods adopt a *Parallel Coupling* strategy, where the physical and data-driven branches operate as independent, concurrent predictors. However, our ablation study reveals a severe performance drop when reverting to this architecture.

The core issue lies in **functional redundancy**. In a parallel design, both branches attempt to model the dominant macroscopic transport trends. Since deep neural networks possess massive parameter capacity, the data-driven branch tends to "lazily" rediscover the basic diffusion-advection laws already captured by the physical ODE. During joint backpropagation, this overlap induces **gradient competition**, where the data-driven loss dominates the optimization landscape, effectively overriding and marginalizing the explicit physical prior.

We further examine gradient interactions and branch specialization. Let $\mathbf{g}^{P}_{\mathrm{pred}} = \nabla_{\Theta_{\mathrm{phy}}} \mathcal{L}_{\mathrm{pred}}$ and $\mathbf{g}_{\mathrm{phy}} = \nabla_{\Theta_{\mathrm{phy}}} \mathcal{L}_{\mathrm{phy}}$. Table 12 summarizes the results.

Table 12: Coupling diagnostics for parallel coupling and PAS^2. Sudden-change MAE is reported on KnowAir.

Diagnostic	Parallel	PAS^2
Negative-cosine steps (%)	61.8	N/A
Gradient-norm ratio	0.92	0.00
Branch-increment correlation	0.72	0.19
Neural–residual correlation	0.49	0.82
Sudden-change MAE	54.76	48.41

Under parallel coupling, the two physical-branch gradients frequently point in opposing directions. In PAS^2, gradient detachment removes the prediction-loss gradient to the physical parameters, making their cosine similarity undefined. The lower branch-increment correlation and higher neural–residual correlation are consistent with reduced redundancy.

In contrast, our Physics-Anchored Decomposition avoids this by decoupling the target. By explicitly directing the neural branch to learn the *residual dynamics* ($\hat{X}^R$), we constrain the network to focus solely on the high-frequency, non-stationary signals that the physical model inherently fails to capture.

This guarantees that neural capacity is efficiently allocated and encourages complementary residual learning.

P.2 The Low-Pass Dilemma and Spectral Transients (S^2 Reasoning)

The ablation of the S^2 Reasoner directly targets the Low-Pass Dilemma prevalent in standard Graph Neural Networks.

Failure of *w/o Spectral Path*: GNN aggregators (the Spatial Path) inherently function as low-pass filters; by aggregating information from neighbors, they inevitably average out high-frequency spatial transients. The results on the *Sudden Change* subset empirically confirm this limitation: without the Spectral Path, the model's MAE and RMSE spike dramatically when extreme pollution peaks occur. As shown in Fig. 21 The spectral features successfully recover these transients because high-frequency anomalies manifest as distinct, isolated modes in the frequency domain, making them easily distinguishable by our Frequency Transformer.

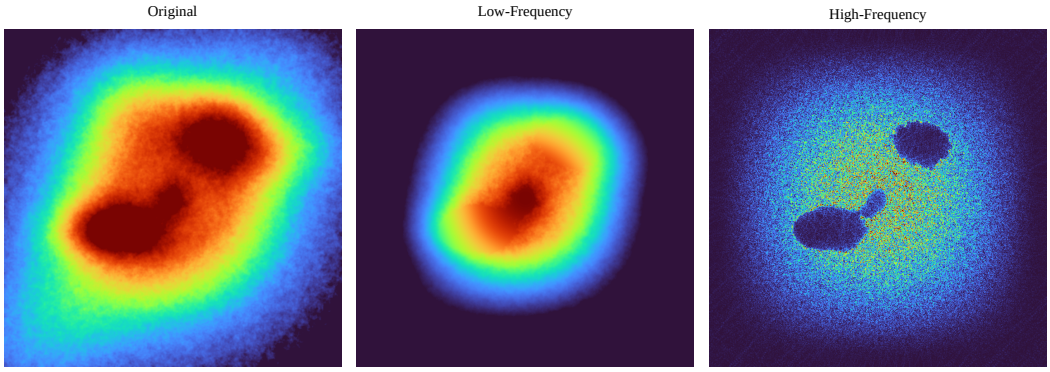


Figure 21: Concentration characteristics of air pollutants at low and high frequencies.

Fourier vs. Wavelet. We replace the FFT/IFFT with a parameter-matched two-level Haar DWT/IDWT, keeping the remaining architecture and training settings unchanged. Table 13 compares the two variants on sudden-change subsets.

Table 13: Fourier and Haar wavelet comparison on sudden-change subsets (MAE/RMSE).

Variant	KnowAir	Beijing
Haar Wavelet	49.58 / 70.16	151.41 / 221.72
Fourier (PAS^2)	**48.41 / 68.01**	**148.46 / 217.80**

Both variants improve over the model without the Spectral Path. Fourier achieves lower sudden-change errors on both datasets, supporting its use alongside the locality-preserving Spatial Path.

Comparison with Non-Spectral Global Reasoning. To examine whether the gains arise from frequency-domain reasoning rather than model capacity or global attention alone, we replace the Spectral Transformer with parameter-matched Global Attention or Graph Transformer modules. All variants use the same data splits, hidden dimensions, and training settings. Table 14 reports the results.

Table 14: Comparison with parameter-matched non-spectral alternatives. Results are 3-day/sudden-change MAE.

Variant	KnowAir	Beijing
Global Attention	19.13 / 50.76	32.58 / 154.89
Graph Transformer	19.24 / 51.10	32.75 / 155.61
PAS^2	**18.70 / 48.41**	**31.24 / 148.46**

PAS^2 achieves lower errors than both alternatives, particularly on sudden-change samples. This suggests that frequency-domain interactions offer benefits beyond matched model capacity and global reasoning alone.

Failure of *w/o Spatial Path* and *w/o Dual-Domain Gating*: Conversely, discarding the Spatial Path degrades overall 3-day accuracy, highlighting that smooth diffusion modeling is indispensable for baseline prediction. Furthermore, the drop in *w/o Dual-Domain Gating* (using a fixed mixing rule instead of learnable attention) validates that the balance between spectral sharpness and spatial smoothness is non-stationary and must be dynamically calibrated based on the local physical context.

P.3 Importance of Sparsity and Physical Guidance in Regime Selection

Our Regime-Adaptive Encoder addresses the multi-modal nature of atmospheric transport. The ablation results validate two critical design choices:

1) **The Necessity of Expert Sparsity:** The *w/o Expert Sparsity* variant activates all attention heads densely. This results in a "feature dilution" effect, where the distinct signals of extreme weather experts are averaged out by the inactive background experts. Enforcing Top-m sparsity ensures the extracted context remains sharp and regime-specific. 2) **Physics-Conditioned Routing:** Removing the physical context from the gating mechanism (*w/o Physics-Conditioned Routing*) causes a notable performance drop, especially in sudden changes. This corroborates that purely data-driven routers struggle to identify the correct regime transition without explicit physical cues.

P.4 Constraining the Solution Space with Physical Context

Finally, the variants *w/o Physics Context Injection* and *w/o Physics Context Encoder* validate our physical conditioning hypothesis. Neural networks are unconstrained universal function approximators; without proper grounding, they may learn mapping functions that violate physical causality.

By injecting the physical context vector C^P (which encodes boundary fluxes and the diffusion-advection ratio α) into the neural Reasoner, we effectively impose a dynamic boundary condition. This explicit context injection constrains the high-dimensional solution space, ensuring that the learned residual corrections are strictly inextricably linked to the underlying physical states, rather than unconstrained, statistically spurious mappings.

Q Visualization of Sudden Changes

To qualitatively validate the capability of the PAS^2 framework in handling extreme non-stationary dynamics, we visualize the prediction trajectories during Sudden Change events at specific monitoring stations.

Existing state-of-the-art models, such as Air-DualODE, often capture the overall upward or downward trends but fail to accurately predict the peak pollutant concentrations. This limitation arises from the inherent "Low-Pass Dilemma" of standard spatial aggregators, which tend to oversmooth high-frequency signals. In contrast, as illustrated in Fig. 22, PAS^2 successfully overcomes this bottleneck.

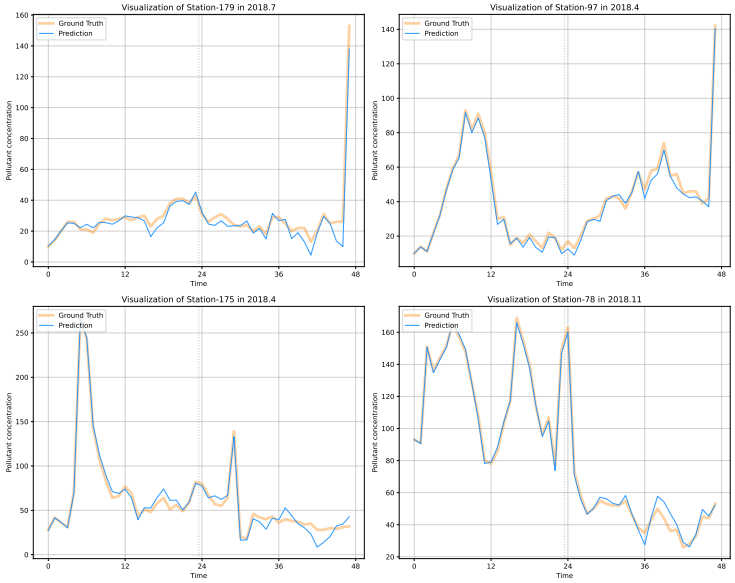


Figure 22: Visualization of sudden change predictions on the KnowAir dataset. Panels (a) and (b) illustrate the model's ability to capture abrupt upward accumulation trends, while panels (c) and (d) demonstrate accurate forecasting of rapid downward dissipation. In all cases, PAS^2 successfully reaches the extreme peak magnitudes typically oversmoothed by baseline models.

The visualization demonstrates the model's response to abrupt pollution accumulation (upward trends) and rapid dissipation events (downward trends) driven by strong ventilation. Notably, PAS^2 not only aligns with the temporal phase of the changes but also accurately reaches the extreme peak magnitudes without temporal lag.

These results empirically corroborate the efficacy of our proposed S^2 Reasoner. By explicitly modeling spectral transients in the frequency domain, PAS^2 recovers the fine-grained high-frequency dynamics that are typically lost in purely spatial paradigms, thereby providing highly reliable early-warning support for downstream environmental policy-making.

R Additional Experimental Results

To further substantiate the long-term robustness of the PAS^2 framework, we decompose the 72-hour forecasting horizon into three fine-grained 24-hour intervals (1–24h, 24–48h, and 48–72h). We first present the temporal degradation evaluation on the **KnowAir** benchmark in Table 15.

Table 15: Fine-grained temporal performance comparison on the **KnowAir** dataset across three 24-hour intervals. The best results are highlighted in red and the second-best in purple.

Methods	1-24h			24-48h			48-72h		
	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
HA	29.43	39.21	0.61	29.41	39.19	0.61	29.38	39.14	0.61
VAR	27.33	36.16	0.50	27.21	35.70	0.61	28.46	37.83	0.64
LatentODE	21.85	30.74	0.49	23.33	33.29	0.51	23.85	34.18	0.52
ODE-RNN	20.93	30.17	0.49	22.94	33.04	0.50	23.90	34.24	0.52
ODE-LSTM	16.86	25.85	0.38	25.48	35.57	0.52	29.07	40.06	0.57
DCRNN	23.15	34.13	0.52	23.47	35.50	0.53	24.02	36.62	0.54
STGCN	22.09	31.32	0.48	24.43	35.87	0.51	26.00	37.66	0.54
ASTGCN	15.67	25.39	0.36	20.70	31.98	0.45	21.26	34.83	0.47
GTS	23.85	33.37	0.53	24.19	34.68	0.54	25.38	36.56	0.56
GMAN	18.32	29.44	0.41	20.80	32.67	0.46	21.79	34.09	0.48

Continued on next page...

Table 15 continued from previous page

Methods	1-24h			24-48h			48-72h		
	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
STTN	18.09	27.52	0.43	27.72	40.13	0.57	33.43	48.35	0.63
$PM_{2.5}$-GNN	20.88	31.39	0.51	23.06	34.55	0.52	23.48	35.08	0.53
AirFormer	15.66	25.70	0.36	19.97	31.47	0.46	21.04	33.34	0.49
AirPhyNet	17.70	28.46	0.40	21.23	32.84	0.47	22.64	34.79	0.49
Air-DualODE	15.59	25.83	0.36	19.89	31.50	0.46	21.70	33.03	0.48
PAS^2	15.48	25.27	0.36	19.74	31.23	0.44	20.88	32.55	0.46

Due to the inherent compounding of epistemic and aleatoric uncertainties in dynamic atmospheric systems, all baseline models exhibit a monotonic decline in predictive accuracy as the horizon extends. However, as shown above, PAS^2 consistently demonstrates superior temporal stability. This sustained robustness is further corroborated by the results on the **Beijing** dataset, detailed in Table 16.

Table 16: Fine-grained temporal performance comparison on the **Beijing** dataset. Format strictly follows Table 15.

Methods	1-24h			24-48h			48-72h		
	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
HA	59.17	87.71	0.82	58.88	87.51	0.82	58.59	87.31	0.81
VAR	45.93	85.63	0.88	78.21	135.31	1.28	80.21	138.91	1.28
LatentODE	36.16	80.22	0.55	38.30	84.55	0.57	40.83	87.29	0.60
ODE-RNN	40.02	76.79	0.63	44.00	80.97	0.67	46.78	83.24	0.71
ODE-LSTM	44.27	75.40	0.65	66.87	95.23	0.88	74.47	101.29	0.92
DCRNN	34.69	67.33	0.68	43.67	84.71	0.78	55.09	96.72	0.85
STGCN	33.41	65.36	0.58	36.46	73.66	0.57	40.56	81.03	0.59
ASTGCN	26.48	59.98	0.46	39.53	81.50	0.61	46.34	89.30	0.67
GTS	38.34	68.79	0.62	53.23	88.17	0.81	58.56	97.54	0.88
GMAN	58.72	92.95	0.81	57.73	91.73	0.79	56.42	89.23	0.77
STTN	44.64	70.35	0.66	72.85	101.83	0.88	75.94	95.01	0.80
$PM_{2.5}$-GNN	28.08	63.93	0.48	40.53	85.39	0.62	45.77	90.42	0.67
AirFormer	25.63	54.83	0.50	36.64	74.96	0.54	39.63	80.92	0.57
AirPhyNet	23.72	56.24	0.48	33.20	74.24	0.52	39.47	84.96	0.58
Air-DualODE	38.41	81.42	0.58	37.80	85.31	0.56	38.85	81.41	0.59
PAS^2	23.56	53.91	0.44	32.76	71.62	0.52	37.41	79.33	0.56

Critically, in the most challenging 48–72h interval—where purely data-driven architectures (e.g., ASTGCN, AirFormer) and decoupled Neural ODEs suffer from severe performance degradation—PAS^2 exhibits the lowest deterioration rate and maintains state-of-the-art accuracy across both datasets. This sustained performance empirically validates that our Physics-Anchored Decomposition mechanism successfully grounds the forecasting trajectory within deterministic transport laws, effectively mitigating error accumulation over extended horizons.

R.1 Evaluation on the CCAQ Benchmark

We further evaluate PAS^2 on CCAQ, an additional nationwide air-quality benchmark, following its standard evaluation protocol. Table 17 compares the forecasting MAE with AirFormer and AirPhyNet.

PAS^2 achieves the lowest MAE among the compared models on CCAQ. This result provides additional evidence of its forecasting performance on a third monitoring benchmark.

R.2 Generalization to Unseen Monitoring Stations

We evaluate spatial generalization using five-fold station holdout experiments. In each fold, 20% of monitoring stations are excluded from the training and validation graphs. At inference, predictions for

Table 17: Forecasting performance on the CCAQ benchmark (MAE; lower is better).

Model	MAE
AirFormer	18.16
AirPhyNet	17.94
PAS^2	**16.52**

these unseen stations use their historical observation windows and geographic coordinates. Table 18 reports the forecasting performance.

Table 18: Forecasting performance at unseen monitoring stations (MAE; lower is better).

Dataset	Model	3-day	Sudden Change
KnowAir	AirFormer	22.68	57.41
	PAS^2	**21.79**	**54.22**
Beijing	AirPhyNet	37.46	168.90
	PAS^2	**35.94**	**162.37**

PAS^2 achieves lower errors than the strongest reported baseline on both datasets. These results support its ability to generalize to monitoring locations excluded from model fitting.

R.3 Evaluation across Chronological Periods

We further examine forecasting performance across four disjoint chronological periods of the Beijing sudden-change subset. Table 19 reports the sudden-change MAE of PAS^2 and the strongest baseline in each period.

Table 19: Sudden-change MAE across four disjoint Beijing periods (lower is better). The strongest baseline is selected separately for each period.

Period	PAS^2	Best Baseline	Baseline Model
1	**154.82**	159.30	AirPhyNet
2	**160.37**	162.29	AirFormer
3	**151.94**	157.12	STTN
4	**146.11**	147.23	AirPhyNet
Mean $\pm$ Std.	**153.31** $\pm$ **5.94**	156.49 $\pm$ 6.52	–

PAS^2 achieves lower sudden-change MAE than the strongest baseline in all four periods. These results indicate consistent forecasting performance across the evaluated time periods.

R.4 Statistical Significance and Practical Relevance

We evaluate the consistency of the forecasting improvements across random seeds and disjoint chronological periods. As reported in Table 8, five independent runs with fixed data splits yield a sudden-change MAE of 148.46 ± 0.92 for PAS^2 and 152.14 ± 1.02 for AirPhyNet on Beijing. These results show stable performance across different random initializations and mini-batch sampling orders.

We further examine temporal consistency using the four disjoint Beijing periods in Table 19. PAS^2 achieves the lowest sudden-change MAE in every period, although the strongest competing model varies across periods. For statistical testing, we use AirPhyNet as a fixed comparator and conduct a two-sided paired t-test on the period-wise MAE differences. The mean MAE reduction is approximately 3.61, with a 95% confidence interval of $[0.13, 7.09]$ ($p = 0.046$). The positive interval supports an average improvement in this comparison, while its width reflects uncertainty in the magnitude of the gain.

To complement the MAE-based analysis with a task-level evaluation, we further examine warning-oriented event detection. As detailed in Appendix R.5, PAS^2 achieves higher event-level F1 and

lower false-negative rates on both datasets, indicating that the improvements in sudden-change forecasting also extend to fewer missed high-concentration transitions.

R.5 Warning-Oriented Event Detection

Beyond regression errors on sudden-change samples, we evaluate the ability of forecasting models to identify warning-relevant pollution events. Following the sudden-change definition in Appendix F, a positive event corresponds to a station–time transition where the $PM_{2.5}$ concentration exceeds $75\,\mu g/m^3$ and the absolute concentration change exceeds $20\,\mu g/m^3$ within a three-hour interval. We apply the same concentration and change-rate thresholds to the predicted and ground-truth sequences to obtain event labels.

We report event-level F1 and false-negative rate (FNR):

$$\mathrm{F1} = \frac{2TP}{2TP + FP + FN}, \qquad \mathrm{FNR} = \frac{FN}{TP + FN}, \tag{63}$$

where TP, FP, and FN denote true-positive, false-positive, and false-negative station–time events, respectively. F1 evaluates the balance between event precision and recall, while FNR measures the proportion of observed events missed by the predicted event labels.

Table 20: Warning-oriented event detection on KnowAir and Beijing. Higher F1 and lower FNR are better.

Dataset	Model	F1	FNR
KnowAir	AirFormer	0.691	0.302
	PAS^2	**0.737**	**0.249**
Beijing	AirPhyNet	0.663	0.326
	PAS^2	**0.705**	**0.281**

PAS^2 achieves higher event-level F1 and lower FNR on both datasets. Compared with AirFormer on KnowAir, F1 increases by 0.046 and FNR decreases by 0.053. On Beijing, the corresponding improvements over AirPhyNet are 0.042 and 0.045. These results complement concentration-level MAE and RMSE by showing that improved forecasting accuracy is accompanied by better identification of abrupt, high-concentration pollution events, including fewer missed events.

S Computational Efficiency and Scalability

S.1 Scalability to Network Size

Scalability is a critical factor for city-wide deployment. We simulated the inference time scaling with respect to the number of monitoring stations N. As illustrated in Figure 23, traditional Transformer-based models exhibit $\mathcal{O}(N^2)$ complexity due to global self-attention. In contrast, PAS2 maintains near-linear scaling $\mathcal{O}(N)$ thanks to the sparse graph topology and efficient ODE solvers, making it suitable for large-scale networks (e.g., $N > 500$).

S.2 Training Efficiency and Memory

Figure 24 compares the validation loss convergence speed. PAS2 achieves competitive performance in significantly fewer wall-clock minutes compared to recurrent baselines, attributed to the parallelizable nature of the ODE solver on GPUs. Furthermore, Figure 25 demonstrates that our model has a smaller memory footprint, allowing for larger batch sizes or longer historical windows.

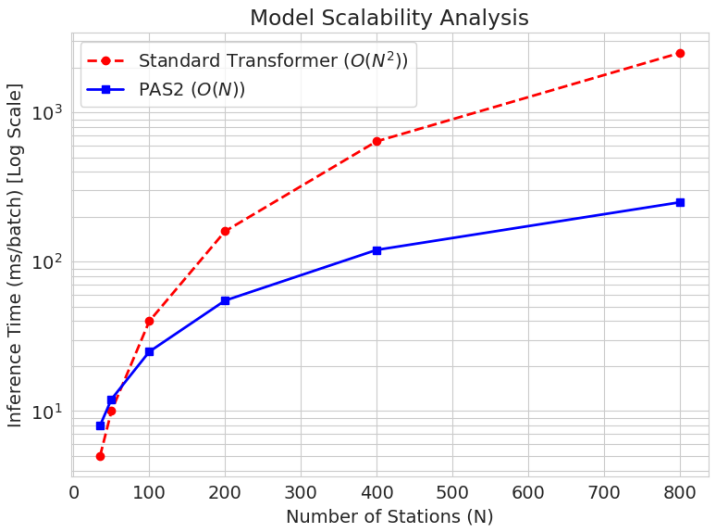


Figure 23: Inference time scalability w.r.t the number of stations. PAS2 shows linear scaling characteristics.

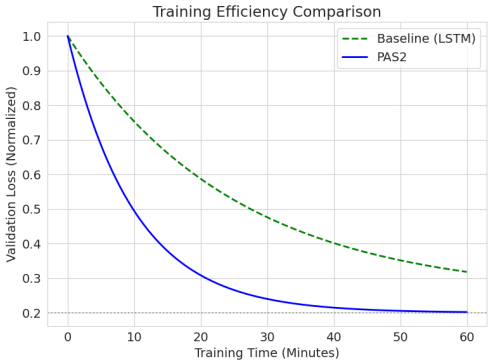


Figure 24: Training Convergence Time.

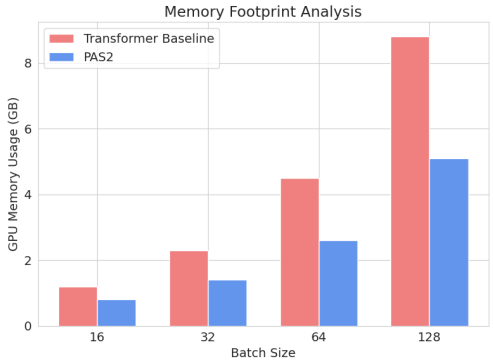


Figure 25: GPU Memory Consumption.

S.3 Forecasting Stability

A key advantage of continuous-time modeling is robustness over long horizons. Figure 26 shows the error accumulation over a 48-hour forecast window. Conventional RNNs suffer from rapid error propagation. The ODE-based dynamics of PAS2 ensure a stable trajectory, significantly reducing error in long-term predictions.

T Implementation Details

T.1 Station-to-Grid Projection

The projection uses geographical coordinates rather than station indices or ordering. For each dataset, station coordinates are normalized to $[0, 1]^2$ within a fixed geographical bounding box. We project

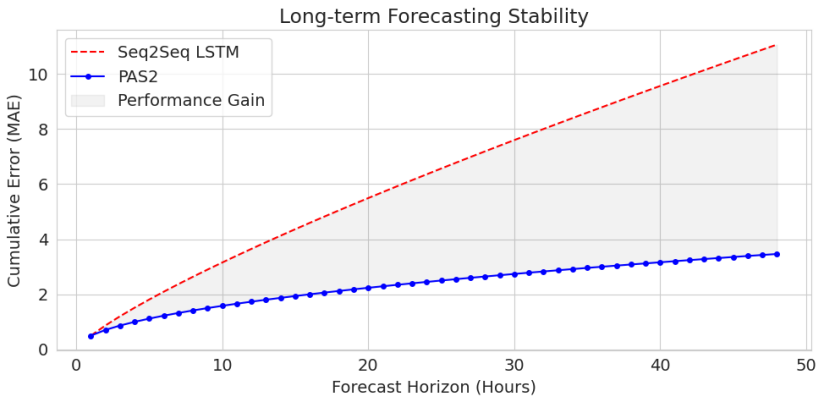


Figure 26: Comparison of error accumulation over increasing forecast horizons (1-48 hours).

station features onto a 32×32 grid using four-nearest-neighbor inverse-distance weighting:

$$\mathbf{U}_{hw} = \frac{\sum_{i \in \mathcal{N}_4(\mathbf{g}_{hw})} w_{i,hw} \mathbf{z}_i}{\sum_{i \in \mathcal{N}_4(\mathbf{g}_{hw})} w_{i,hw}}, \qquad w_{i,hw} = \left(\|\mathbf{g}_{hw} - \tilde{\mathbf{p}}_i\|_2 + 10^{-8} \right)^{-2}. \tag{64}$$

Here, $\mathbf{g}_{hw}$ is a grid location, $\tilde{\mathbf{p}}_i$ is the normalized station coordinate, and $\mathbf{z}_i$ is its feature vector.

Missing observations are imputed before projection, as described in Appendix F. No additional mask or padding is applied to the interpolated FFT grid. The DFT uses its standard periodic boundary convention.

T.2 Software Environment

The experiments were conducted on a Linux server with the software environment detailed in Table 21.

Table 21: Software Environment and Dependencies

Library/Tool	Version
Python	3.9.7
PyTorch	1.10.1
NumPy	1.21.2
Torchdiffeq	0.2.2
CUDA	11.3

U Robustness and Error Analysis

U.1 Robustness to Meteorological Input Noise

We examine forecasting robustness under noisy historical meteorological observations. At inference time, we add zero-mean Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the z-score-normalized meteorological inputs, with $\sigma \in \{0.1, 0.2\}$. All models are evaluated without retraining under the same perturbation protocol. Table 22 reports the 3-day MAE of PAS^2, the physics-only predictor, and the strongest competing baseline under each setting.

Forecasting errors increase as the meteorological inputs become noisier. However, PAS^2 shows smaller error increases than the physics-only predictor and maintains the lowest MAE among the evaluated models under both noise levels. These results indicate that residual correction helps maintain forecasting accuracy when meteorological observations are perturbed.

Table 22: Robustness to noisy historical meteorological inputs (3-day MAE; lower is better). The strongest competing baseline is identified separately for each setting.

Dataset	Noise level	Physics-only	Best Baseline	PAS^2
KnowAir	Clean	24.88	18.89 (AirFormer)	**18.70**
	$\sigma = 0.1$	26.36	19.61 (Air-DualODE)	**19.26**
	$\sigma = 0.2$	28.71	20.68 (ASTGCN)	**20.13**
Beijing	Clean	45.36	32.13 (AirPhyNet)	**31.24**
	$\sigma = 0.1$	48.11	33.37 (AirPhyNet)	**32.18**
	$\sigma = 0.2$	52.14	35.16 (Air-DualODE)	**33.74**

Table 23: Sensitivity to diffusion–advection operator misspecification (3-day MAE; lower is better). Results under perturbation are averaged over the two scaling directions.

Setting	KnowAir		Beijing	
	Physics-only	PAS^2	Physics-only	PAS^2
Clean	24.88	**18.70**	45.36	**31.24**
Operator $\pm 10\%$	26.41	**19.18**	48.27	**32.09**
Operator $\pm 20\%$	28.37	**19.85**	51.76	**33.12**

U.2 Robustness to Physical Operator Misspecification

We further investigate the sensitivity of PAS^2 to misspecification of the physical transport operators. At inference time, we perturb the diffusion and advection strengths by multiplicative factors of 0.9 or 1.1 ($\pm 10\%$) and 0.8 or 1.2 ($\pm 20\%$), without retraining. For each perturbation level, the reported errors are averaged over the two perturbation directions. We compare the physics-only predictor with the complete PAS^2 framework to examine how residual correction responds to deviations in the physical anchor.

Both predictors exhibit increased errors under stronger operator perturbations, but the complete PAS^2 framework shows substantially smaller degradation. Under the $\pm 20\%$ setting, the physics-only MAE increases by 3.49 on KnowAir and 6.40 on Beijing, whereas the corresponding increases for PAS^2 are 1.15 and 1.88. This difference indicates that the learned residual branch can compensate for part of the forecasting error introduced by misspecified physical transport strengths.

U.3 Error Distribution

Figure 27(a) displays the histogram of prediction residuals ($y_{true} - y_{pred}$). The deviations are normally distributed around zero, suggesting the model is unbiased. Figure 27(b) confirms the model's accuracy across different concentration levels.

U.4 Spatial Performance Variance

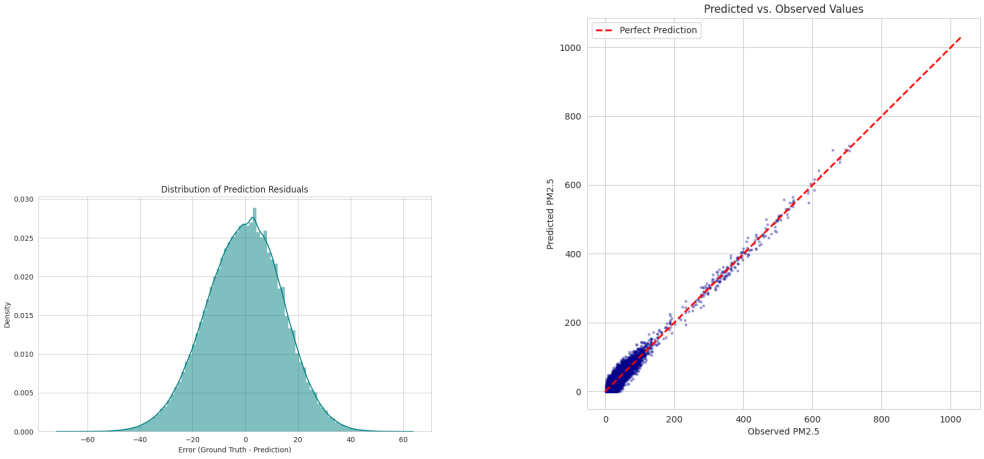
Figure 28 presents the RMSE performance across the 35 monitoring stations.

V Hyperparameter Sensitivity Landscape

The PAS^2 framework involves two critical hyperparameters governing the physics-data synergy: the physics supervision weight λ_1 and the open-boundary regularization weight λ_2. We conduct a grid search to analyze the impact of these coefficients on the model's performance, using RMSE as the primary validation metric.

Impact of Physical Supervision (λ_1). The coefficient λ_1 controls the strength of the physical supervision loss imposed on the physical anchor.

- **Small λ_1 (< 0.1):** The physical estimator receives weak supervision and the model tends to behave closer to a purely data-driven predictor. In this case, the neural branch may absorb



(a) Residual Distribution

(b) Prediction vs. Ground Truth

Figure 27: Error Analysis of PAS2.

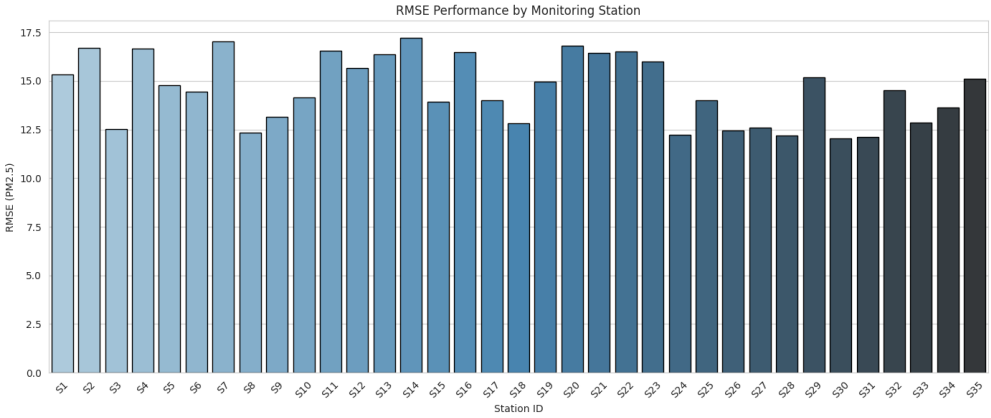


Figure 28: Model Performance (RMSE) across individual monitoring stations.

both macroscopic transport trends and residual fluctuations, weakening the intended role separation between the physical anchor and the neural residual branch.

- **Large λ_1 (> 10.0):** The optimization becomes overly constrained by the simplified physical anchor, limiting the neural branch's ability to correct residual errors that are not explained by diffusion, advection, or open-boundary source–sink effects.
- **Moderate λ_1:** The best validation performance is consistently achieved when λ_1 lies in a moderate range, where the physical branch is sufficiently supervised while the residual branch retains enough flexibility to model non-stationary fluctuations.

Impact of Open-Boundary Regularization (λ_2). The coefficient λ_2 controls the strength of the regularization term on the open-boundary source–sink coefficient β. This term prevents the physical anchor from explaining all deviations through unconstrained source–sink compensation.

- **Small λ_2 ($\lambda_2 \approx 0$):** When the regularization is too weak, the learned source–sink coefficient β becomes overly flexible. Although this may reduce training error, it can weaken the physical meaning of the anchor by allowing external source–sink terms to absorb residual fluctuations that should instead be handled by the neural branch. This leads to less stable validation performance, especially under sudden-change scenarios.

- **Large λ_2:** When λ_2 is too large, the source–sink coefficient is excessively suppressed. The physical anchor then behaves closer to a closed diffusion–advection system and becomes less capable of representing emissions, deposition, regional inflow, and other open-boundary effects. As a result, the physical baseline becomes under-expressive and the residual branch must compensate for larger systematic errors.

- **Moderate λ_2:** A moderate value of λ_2 achieves the best trade-off between physical flexibility and regularization. It allows β to capture necessary open-boundary effects while preventing implausible source–sink compensation. This supports the design choice of using $\lambda_2 \mathcal{L}_{\mathrm{reg}}$ to regularize the physical estimator rather than leaving the open-boundary term unconstrained.

Overall, the sensitivity analysis shows that both λ_1 and λ_2 need to be balanced. The former determines how strongly the physical anchor is supervised, while the latter controls the flexibility of the open-boundary source–sink term.

Sensitivity to Grid Resolution and Interpolation. We evaluate three grid resolutions and different station-to-grid interpolation methods. Each variant is retrained using the same data splits and training settings. The default configuration uses 32×32 IDW, while the nearest-neighbor and RBF variants retain the same grid resolution. Table 24 reports the forecasting errors.

Table 24: Sensitivity to grid resolution and interpolation (3-day/sudden-change MAE).

Projection setting	KnowAir	Beijing
16×16 IDW	19.01 / 49.35	31.83 / 150.72
32×32 IDW (default)	**18.70 / 48.41**	**31.24 / 148.46**
64×64 IDW	18.89 / 48.96	31.58 / 149.64
32×32 Nearest-neighbor	19.16 / 50.02	32.05 / 152.31
32×32 RBF	18.82 / 48.77	31.39 / 149.02

The results show modest performance variations across the tested resolutions and interpolation methods. The default 32×32 IDW configuration achieves the lowest errors on both datasets.

Sensitivity to Station Density. We further evaluate the effect of reduced station coverage during inference. We retain 80% or 60% of the stations for grid projection without retraining the model. The graph branch and evaluation targets remain unchanged, and results are averaged over five station masks.

Table 25: Sensitivity to station density during inference (3-day/sudden-change MAE).

Stations retained	KnowAir	Beijing
100% (default)	18.70 / 48.41	31.24 / 148.46
80%	19.06 / 49.52	31.86 / 151.10
60%	19.52 / 51.03	32.74 / 155.36

Forecasting errors increase as fewer stations contribute to the grid projection, but the degradation remains moderate under the tested settings.

Sensitivity to Geospatial Graph Construction. We examine the sensitivity of PAS^2 to the elevation-barrier threshold τ_h and the initialization of the learnable RBF bandwidth σ_0. The elevation threshold controls the topographic connectivity mask, while the RBF bandwidth determines the distance-dependent weights of the diffusion graph. We vary one parameter at a time and retrain each configuration under the same experimental settings. Table 26 reports the 3-day and sudden-change forecasting errors on Beijing.

The default configuration achieves the lowest forecasting errors among the tested settings. Relative to the default, the 3-day MAE increases by at most 0.72, while the sudden-change MAE increases by at most 2.72. These results indicate limited sensitivity to the tested variations in topographic connectivity and diffusion-kernel initialization.

Table 26: Sensitivity to graph construction parameters on Beijing (MAE; lower is better).

Setting	3-day MAE	Sudden-change MAE
Default	**31.24**	**148.46**
Elevation threshold $0.8\tau_h$	31.72	150.31
Elevation threshold $1.2\tau_h$	31.55	149.57
RBF initialization $0.5\sigma_0$	31.96	151.18
RBF initialization $2.0\sigma_0$	31.79	150.52

W Limitations

Our work currently has several limitations. First, although the stop-gradient design structurally separates the physical anchor from the neural residual branch and reduces direct cross-branch interference, it should not be interpreted as a general proof of conflict-free multi-objective optimization. The physics supervision and open-boundary regularization still operate within the physical parameter subset, and a formal guarantee for all within-branch gradient interactions remains an open direction. Second, the current physical anchor models diffusion, advection, and an effective source–sink contribution, but does not explicitly resolve chemical transformations (e.g., secondary aerosol formation), deposition processes, or detailed emission inventories. These unresolved processes may influence the learned source–sink term and neural residual, but their individual contributions cannot be identified from either component. Consequently, the physical anchor remains a simplified representation of atmospheric pollutant dynamics and may be affected by structural misspecification when these processes or boundary conditions vary. Incorporating reaction–transport dynamics, explicit deposition, and emission information could improve its physical fidelity and interpretability. Third, the perturbation experiments in Appendix U assess the sensitivity of forecasting accuracy to noisy meteorological inputs and misspecified physical transport strengths, but do not establish predictive uncertainty calibration. The current framework produces point forecasts without calibrated prediction intervals or probabilities of exceeding pollution thresholds. Extending PAS^2 with probabilistic forecasts and evaluating their empirical coverage and threshold-event calibration would be valuable for risk-sensitive air-quality forecasting. Finally, our experiments focus on $PM_{2.5}$ forecasting using KnowAir, Beijing, and CCAQ, all drawn from monitoring networks in China. The nationwide benchmark evaluations and five-fold station-holdout experiments extend the spatial scope of the evaluation, but transfer to networks in other countries or held-out climatic regimes remains to be investigated. The current framework has also not been evaluated for other pollutant species or different spatiotemporal forecasting tasks. Such extensions would require appropriate physical operators and further empirical validation.

X Broader Impacts

This paper aims to advance machine learning methods for Earth system modeling, with a focus on air quality forecasting. The proposed PAS^2 framework can improve forecasting accuracy, particularly during high-volatility "Sudden Change" events. Its potential societal benefits include more reliable early pollution warnings, better support for public health protection, and more informed environmental policy-making.

At the same time, inaccurate forecasts during extreme pollution episodes may have significant downstream consequences. Underprediction or missed pollution events may delay warnings and protective actions, increasing exposure risks, particularly for vulnerable populations. Conversely, false alarms may contribute to alert fatigue, unnecessary restrictions, and misallocation of public resources. Therefore, PAS^2 should serve as a decision-support tool rather than an autonomous trigger for public-health or regulatory actions. Practical deployment would require local validation, calibrated warning thresholds, uncertainty assessment, human expert review, and conservative escalation procedures near hazardous pollution levels.

Y Discussion

Y.1 Significance and Real-World Implications

The results presented in Section 4.2 demonstrate that the proposed PAS^2 strategy consistently outperforms existing state-of-the-art baselines across varying time horizons and spatial scales. Notably, the model excels in capturing extreme non-stationary dynamics, showcasing a significant reduction in prediction errors during “Sudden Change" events.

These results also highlight the potential value of PAS^2 for pollution warning and environmental decision support. Beyond improvements in concentration forecasting, the event-level evaluation in Appendix R.5 shows higher F1 scores and lower false-negative rates on both datasets, indicating better identification of abrupt, high-concentration pollution events under the evaluated warning thresholds. The station-holdout experiments in Appendix R.2 further demonstrate predictive transfer to monitoring locations excluded from model fitting, using their available historical observations and geographic coordinates. Together, these findings motivate further investigation of warning-oriented applications, where locally validated thresholds, uncertainty assessment, and expert oversight are important for reliable operational use.

Y.2 Architectural Advantages over Existing Physics-Guided Models

To better understand the superiority of PAS^2, we provide a structural comparison with two leading Physics-Guided Learning (PGL) methods: AirPhyNet (Hettige et al., 2024) and Air-DualODE (Tian et al., 2025), summarized in Table 27.

Resolving Functional Redundancy. Both AirPhyNet and Air-DualODE adopt a “Parallel Coupling" paradigm, where the physical and neural branches operate as peer predictors. This configuration inadvertently induces gradient competition, often causing the high-capacity neural branch to overshadow the physical prior. In contrast, PAS^2 introduces a *Physics-Anchored* paradigm. By strictly constraining the neural branch to learn the complementary residual, we eliminate functional redundancy and ensure true synergy.

Overcoming the Low-Pass Dilemma. While AirPhyNet and Air-DualODE rely heavily on spatial graph operators (which act as low-pass filters), PAS^2 pioneers the integration of frequency-domain Spectral Operators in air quality forecasting. This explicitly resolves the “Low-Pass Dilemma," enabling the capture of high-frequency spatial gradients and long-range teleconnections that parallel GNNs naturally oversmooth.

Table 27: Structural comparison between PAS^2 and state-of-the-art physics-guided baselines.

Characteristics	AirPhyNet	Air-DualODE	$\boldsymbol{PAS^2}$ **(Ours)**
System Assumption	Closed System	Open System	Open System
Coupling Paradigm	Parallel Coupling	Parallel Coupling	Anchored Residual Correction
Optimization Conflict	High (Gradient competition)	High (Requires complex alignment)	Low (Strict separation)
Spatial Reasoning	Graph-based ODE (Low-pass)	GNN + Spatial-MSA	S^2
Sudden Change Capability	Weak	Weak	**Strong**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the core contributions and scope of our work. The claims are consistent with the proposed method, theoretical analysis, and empirical results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section that discusses the main limitations of the proposed framework and possible future directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The formal theoretical statements are provided with their assumptions, and the corresponding derivations and proofs are included in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We carefully discuss the details of the experimental implementation in the main text and appendix, and we also provide the code for reference.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the anonymized code of our model for reference, and the datasets used in our experiments are standard public benchmarks with preprocessing details described in the appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide relevant experimental settings in the main text and detailed implementation configurations in the appendix, including data splits, hyperparameters, optimizer, and training protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report mean and standard deviation over five independent randomized trials in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computational environment and resource usage in the appendix, including GPU type, peak memory usage, per-run training time, and estimated total GPU-hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully comply with the NeurIPS Code of Ethics, and our work does not involve human subjects, private data, or high-risk model release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive societal impacts and deployment risks of air quality forecasting in Appendix.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not release high-risk models or scraped datasets, and it poses no such misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite relevant literature throughout the paper and provide the sources of the datasets, baselines, algorithms, training environments, and evaluation metrics in the paper and code.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is available as a zip file.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.