SATCLIP-GNN: COLD-START PM2.5 FORECASTING WITH SATELLITE-DERIVED LOCATION PRIORS

Siam Shibly Antar
School of Computer Science
McGill University
Montreal, Canada
siam.antar@mail.mcgill.ca

Syem Shibly Ador
School of Computer Science
Macquarie University
Sydney, Australia
syemshibly.ador@students.mq.edu.au

Steven H. H. Ding
School of Information Studies
McGill University
Montreal, Canada
steven.h.ding@mcgill.ca

Benjamin C. M. Fung
School of Information Studies
McGill University
Montreal, Canada
ben.fung@mcgill.ca

ABSTRACT

Spatiotemporal $\text{PM}_{2.5}$ forecasting is commonly framed as diffusion on a station graph, yet many strong models rely on learned node embeddings that are tightly coupled to the training station set and can generalize poorly when new monitoring sites are deployed. This paper studies a forecasting pipeline that replaces trainable station embeddings with *frozen SatCLIP* location representations derived from satellite imagery, and combines them with temporal modeling and graph message passing. We additionally investigate a physics-inspired wind-gated dynamic graph enforcing downwind transport.

On a five-year U.S. dataset with 596 stations and a chronological split (train 2019–2021, val 2022, test 2023), a Graph WaveNet baseline provides slightly stronger 2023 performance under a $\text{PM}_{2.5}$-only "proper baseline" setting (24h history → 24h horizon), while our best SatCLIP + static radius-graph variant remains competitive and improves MAE by 32.0% relative to a controlled DCRNN-style proxy baseline. Crucially, under a *spatial holdout* protocol where 20% of stations are excluded from training, SatCLIP improves MAE on *held-out (unseen) stations* from 3.679 (Graph WaveNet) to 3.537, demonstrating a clear cold-start advantage and aligning with the hypothesis that foundation-model priors aid transfer to new sensors.

Training dynamics show that wind-gated connectivity can reach lower *training* PWMSE, yet this does not translate into improved 2023 generalization, suggesting that hard dynamic masking can be brittle when meteorological inputs are noisy or partially imputed. Finally, we include an illustrative extreme-event case study (Station 410250003, $t = 43001$) showing that episode-level behavior can differ from aggregate ranking (Figs. 2–3), motivating evaluation on multiple extreme episodes with stratified metrics.

***Index Terms*—** $\text{PM}_{2.5}$ forecasting, spatiotemporal learning, graph neural networks, foundation models, Environmental AI

1. INTRODUCTION

Fine particulate matter ($\text{PM}_{2.5}$) poses significant public health risks, and timely forecasts can support mitigation during acute episodes, such as wildfire smoke intrusions. Forecasting at monitoring stations is inherently spatiotemporal: concentrations depend on temporal persistence, local emissions, and transport mediated by meteorology.

Many spatiotemporal graph neural networks (ST-GNNs) represent stations as nodes connected by a static or learned adjacency, propagating information via graph convolutions or attention. A practical limitation is that node embeddings are often trained as a lookup table tied to the original station set, complicating *cold-start deployment* when new sensors are added. A second challenge is robustness under distribution shift: correlation-derived connectivity or time-varying graphs can behave unpredictably when meteorology fields are noisy, partially imputed, or at mismatched spatial resolution.

Goal. We study whether *transferable semantic priors* from a geospatial foundation model can replace learned station embeddings and improve generalization to *unseen stations*, and whether a *physics-inspired wind-gated graph* improves or harms performance relative to a stable static graph.

Contributions.

1. **SatCLIP station priors:** We integrate frozen SatCLIP location embeddings as fixed station context, eliminating the need for a trainable station embedding table.
2. **Stronger baselines under a minimal protocol:** We implement Graph WaveNet under the same 24→24 protocol using a recommended $\text{PM}_{2.5}$-only "proper baseline" configuration, and report an optional second run with $\text{PM}_{2.5}$+meteorology.
3. **Cold-start (unseen-station) generalization:** We introduce a spatial holdout protocol (20% stations excluded from training), showing SatCLIP improves MAE on held-out stations by $\sim$3.8% relative to Graph WaveNet (Table 2), supporting foundation-model priors in real deployments.
4. **Static vs. wind-gated connectivity and analysis:** We compare a wind-gated dynamic graph against a stable static radius graph; static connectivity yields the strongest overall SatCLIP performance, while wind gating can reduce *training* PWMSE without improving 2023 generalization. Furthermore, we include an illustrative extreme-event case study and highlight key limitations and follow-ups.

2. RELATED WORK

Spatiotemporal forecasting on sensor networks can be formulated as learning dynamical behavior over a graph, where monitoring stations serve as nodes and edges represent spatial dependencies. DCRNN [1] models this spatiotemporal evolution as diffusion on a fixed graph and is widely used as a reference method for graph-based sequence prediction. Approaches like Graph WaveNet [2] and MTGNN [3] add flexibility by inferring adaptive graph structures but they rely on node embeddings that implicitly bind the model to the specific set of training stations. This dependency becomes especially challenging in *cold-start deployment* scenarios, where new sites come online or existing sites are relocated, and retraining is expensive or impractical. Beyond diffusion- and convolution-based designs (e.g., STGCN [4]), attention-driven ST-GNNs like ASTGCN [5] and adaptive recurrent architectures such as AGCRN [6] further enrich model capacity through learned spatial attention or adaptive graph filters, but generally continue to employ station-specific parameters or embeddings that presuppose a fixed collection of nodes.

At the same time, there is increasing interest in *foundation priors* for Earth observation and geospatial modeling. SatCLIP [7, 8] constructs global location embeddings by aligning geographic coordinates with satellite imagery using contrastive learning, producing representations that encode land cover, urban form, and other semantic characteristics linked to local emission patterns and transport behavior. These priors can be interpreted as a *remote-sensing-derived station context* that does not rely on explicit station identifiers and can thus be reused across different sensor networks. Departing from typical ST-GNN evaluations that operate on a fixed station set, the focus here is on realistic deployment: generating forecasts at stations *omitted during training* while using frozen SatCLIP embeddings as contextual information, leaving the forecasting architecture itself unchanged.

3. PROBLEM SETUP

We forecast a 24-hour $\mathrm{PM}_{2.5}$ trajectory at each station given a 24-hour history window. Let $x_t^{(i)}$ denote $\mathrm{PM}_{2.5}$ at station i and time t. The per-station input vector includes:

$$\mathbf{x}_t^{(i)} = \left[x_t^{(i)},\, u_t^{(i)},\, v_t^{(i)},\, \tau_t^{(i)},\, p_t^{(i)}\right],$$

where u, v are the 10 m wind components (eastward and northward, respectively) from ERA5-Land reanalysis, τ is 2 m near-surface temperature, and p is surface pressure. We use $T_{\text{in}} = 24$ and forecast $T_{\text{out}} = 24$.

4. METHOD

Overview. Our goal is to remove reliance on trainable station-ID embeddings while preserving competitive spatiotemporal modeling. For each station i at forecast time t, the model consumes: (i) a T_{in}=24 hour window of station inputs $\{\mathbf{x}_{t-T_{\text{in}}+1}^{(i)}, \ldots, \mathbf{x}_t^{(i)}\}$ ($\mathrm{PM}_{2.5}$ and meteorology), (ii) a frozen SatCLIP embedding $\mathbf{s}_i$ queried by latitude/longitude as a semantic station prior, and (iii) a station graph $\mathcal{G}_t$ describing spatial interactions. The output is a T_{out}=24 hour $\mathrm{PM}_{2.5}$ forecast trajectory. Figure 1 summarizes the pipeline.

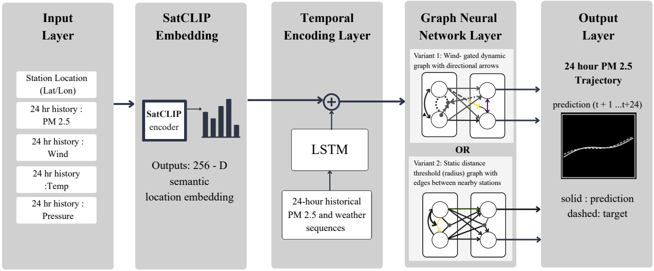
Input Layer
Station Location (Lat/Lon)
24 hr history : PM 2.5
24 hr history : Wind
24 hr history :Temp
24 hr history : Pressure
SatCLIP Embedding
SatCLIP encoder
Outputs: 256 - D semantic location embedding
Temporal Encoding Layer
LSTM
24-hour historical PM 2.5 and weather sequences
Graph Neural Network Layer
Variant 1: Wind- gated dynamic graph with directional arrows
OR
Variant 2: Static distance threshold (radius) graph with edges between nearby stations
Output Layer
24 hour PM 2.5 Trajectory
prediction (t + 1 ...t+24)
solid : prediction
dashed: target

Fig. 1. **SatCLIP-LSTM-GNN for 24h $\mathrm{PM}_{2.5}$ forecasting.** A 2-layer LSTM encodes 24h station history ($\mathrm{PM}_{2.5}$+met), fused with a frozen 256-D SatCLIP location prior; a GNN uses either a static radius or wind-gated graph to predict a 24-step trajectory.

Temporal encoder. A two-layer LSTM summarizes the past window into a station representation $\mathbf{h}_{\text{temp}}^{(i)} \in \mathbb{R}^H$. We use hidden size H=64 with dropout 0.1 to control capacity and reduce overfitting under distribution shift:

$$\mathbf{h}_{\text{temp}}^{(i)} = \mathrm{LSTM}\Big(\mathbf{x}_{t-T_{\text{in}}+1}^{(i)}, \ldots, \mathbf{x}_t^{(i)}\Big)\,, \tag{1}$$

where $\mathbf{x}_t^{(i)} \in \mathbb{R}^F$ contains $\mathrm{PM}_{2.5}$ and meteorology (u, v, τ, p). Overall, this module captures local persistence, daily cycles, and short-term meteorological correlations.

SatCLIP fusion. Each station receives a frozen SatCLIP prior $\mathbf{s}_i \in \mathbb{R}^{256}$ [8]. SatCLIP is trained via contrastive learning on multi-spectral Sentinel-2 satellite imagery and encodes land-surface characteristics—including land cover type, urban density, vegetation fraction, and industrial land use—that correlate with local $\mathrm{PM}_{2.5}$ emission sources and atmospheric dispersion regimes. Compared with generic coordinate encodings (e.g., sinusoidal positional embeddings), SatCLIP captures *semantic* site context derived directly from observed surface conditions, making it particularly suited as a station prior for air quality modeling. A 1-layer MLP $\phi : \mathbb{R}^{256} \to \mathbb{R}^H$ projects $\mathbf{s}_i$ into the temporal hidden space and fuses by residual addition:

$$\mathbf{h}_0^{(i)} = \mathbf{h}_{\text{temp}}^{(i)} + \phi(\mathbf{s}_i). \tag{2}$$

This design ensures (i) the SatCLIP backbone remains frozen, (ii) only a lightweight projection is trained, and (iii) inference at unseen stations is possible as long as coordinates are available.

Spatial message passing. We apply a 2-layer attention GNN (GATv2-style) with 4 heads and hidden size H=64 over node features $\mathbf{H}_0 \in \mathbb{R}^{N \times H}$, producing contextual representations $\mathbf{Z}_t \in \mathbb{R}^{N \times H}$. Message passing captures spatial coupling between stations (transport, shared meteorological regimes, and regional emission patterns). We compare two graph constructions:

(i) Static radius graph: connect stations within r=0.5° (approximately 50 km). This provides a stable spatial prior and avoids time-varying connectivity that can amplify noise.

(ii) Wind-gated dynamic graph: construct a directed, time-varying adjacency by gating edges based on alignment between the local wind vector $\mathbf{w}_t^{(i)} = [u_t^{(i)}, v_t^{(i)}]$ and inter-station displacement $\mathbf{d}_{ij}$:

$$a_{ij}(t) = \frac{\mathbf{w}_t^{(i)} \cdot \mathbf{d}_{ij}}{\|\mathbf{w}_t^{(i)}\|\,\|\mathbf{d}_{ij}\|}, \tag{3}$$

retaining edges when $a_{ij}(t) > \tau$ (optionally with distance decay and a minimum weight ε). This graph is *physics-inspired* in that downwind neighbors should exert stronger influence, but the hard threshold can also make connectivity sensitive to wind noise and temporal mismatch between instantaneous winds and multi-hour transport.

Forecast head and loss. A linear head $\psi : \mathbb{R}^H \to \mathbb{R}^{T_{\text{out}}}$ predicts the 24-step trajectory:

$$\hat{\mathbf{y}}_{t+1:t+T_{\text{out}}}^{(i)} = \psi(\mathbf{z}_t^{(i)}). \tag{4}$$

For high-impact pollution extremes, we use Peak-Weighted MSE:

$$\mathcal{L}_{\text{PWMSE}} = \frac{1}{M}\sum_{m=1}^{M} (y_m - \hat{y}_m)^2\, w_m, \quad w_m = \begin{cases} \lambda & y_m > \mu + \sigma, \\ 1 & \text{otherwise}, \end{cases} \tag{5}$$

where μ, σ are computed from the training $\mathrm{PM}_{2.5}$ distribution and λ=10. This objective increases the penalty on extreme concentrations, reflecting public-health relevance and encouraging sensitivity to episodic events such as wildfire smoke.

5. EXPERIMENTAL SETUP

Dataset and split. We use EPA AQS $\mathrm{PM}_{2.5}$ from 596 U.S. stations (2019–2023) with a chronological split: train 2019–2021, validation 2022, and test 2023. Meteorological inputs are from ERA5-Land reanalysis to provide complementary atmospheric context.

Variants and baselines (24h → 24h). We compare: **SatCLIP-LSTM-GNN (wind-gated)**; **Ablation A (SatCLIP + static radius graph)**; **Ablation B (random priors)**; **LSTM** (temporal-only); **DCRNN proxy** (controlled LSTM+GNN without SatCLIP on a static radius graph); **XGBoost (mean target)**; and **Graph WaveNet** [2] with (i) PM-only inputs under adaptive adjacency, (ii) PM-only with static radius support (r=0.5°), plus optional $\mathrm{PM}_{2.5}$+meteorology runs under both supports.

DCRNN proxy note. The DCRNN proxy is a controlled baseline to isolate the effect of SatCLIP priors and graph construction under a consistent LSTM+GNN backbone; it is not intended as a fully faithful reproduction of all DCRNN implementation details, so we avoid over-claiming against modern ST-GNN baselines.

Training details. Neural models are trained with Adam using the peak-weighted MSE (PWMSE) objective. Unless otherwise stated, normalization statistics are computed from training years only (2019–2021). We report aggregate 2023 test results from checkpoints selected under the same protocol across variants.

Spatial holdout (cold start). We additionally evaluate a station holdout split where 20% of stations are excluded from training and evaluated in 2023. Normalization uses 2019–2021 data restricted to the training-station subset (excluding held-out stations). Performance on held-out stations is reported in Table 2.

6. IMPLEMENTATION DETAILS

Sequence construction. We form supervised samples using a sliding window with T_{in}=24 hours of history to predict T_{out}=24 hours ahead at each station. Targets are the full 24-step $\text{PM}_{2.5}$ trajectory (not a single-step or horizon mean).

Missingness and alignment. EPA AQS measurements contain missing hours; we discard windows with insufficient coverage (or apply light imputation only within the history window), and we align meteorology by matching each station's coordinates to the nearest ERA5-Land grid cell at the corresponding timestamp. All statistics used for scaling (mean/std) are computed on *training years only* (2019–2021), and for station-holdout we compute scaling on the *training-station subset* only.

Graph construction. The static graph uses a radius threshold r=0.5° in lat/long space (approx. 50 km). The wind-gated graph uses directional alignment $a_{ij}(t)$ (Eq. (3)) with threshold τ (and optional distance decay / minimum weight ε) to form a directed, time-varying adjacency. Also, wind gating can induce substantially denser connectivity under some settings, amplifying noise sensitivity.

Training and capacity. All neural variants use Adam with the PWMSE objective (Eq. (5)) and early checkpoint selection under a consistent protocol. The SatCLIP prior is frozen and only a small projection MLP is trained, which reduces reliance on a learned station-ID embedding table and enables cold-start inference for unseen stations without retraining.

Software and data access. The pipeline is implemented in Python using PyTorch and PyTorch Geometric for GNN layers. ERA5-Land reanalysis fields were obtained via the Copernicus Climate Data Store (CDS) API, and hourly $\text{PM}_{2.5}$ observations from the EPA Air Quality System (AQS) Data Mart. Code and configuration files will be released to support reproducibility.

7. RESULTS

Aggregate test performance (2023, seen stations). Table 1 reports MAE/RMSE/R^2 on the standard 2023 test set.

Table 1. 2023 Test (Seen Stations), 24h→24h. GWN= Graph WaveNet; XGB= XGBoost; PM= $\text{PM}_{2.5}$-only inputs; Met= meteorology added; A= adaptive adjacency; S= static radius graph.

Model	MAE	RMSE	R^2
XGB (mean)	2.5205	5.4635	0.5577
LSTM	6.8247	12.0587	-0.4548
DCRNN-proxy	5.3460	11.0843	-0.2292
GWN (PM, A)	3.6287	7.7531	0.3986
GWN (PM, S)	3.6524	7.6387	0.4162
GWN (PM+Met, A)	**3.5811**	7.6642	0.4123
GWN (PM+Met, S)	3.5846	**7.6100**	**0.4206**
Ours (Wind)	5.2438	9.1022	0.1711
Ours (Static)	3.6334	7.6849	0.4091
Ours (Rand)	4.2242	9.2108	0.1512

Seen-station takeaway. Graph WaveNet provides the strongest aggregate 2023 performance among full-horizon (24→24) models, with modest changes when adding meteorology (best R^2=0.4206, best RMSE=7.6100, best MAE=3.5811). Our best SatCLIP variant (Ablation A) remains competitive and reduces MAE from 5.3460 (DCRNN Proxy) to 3.6334 (32.0% reduction).

SatCLIP effect (seen stations). Replacing SatCLIP priors with random vectors increases MAE from 3.6334 to 4.2242 (Ablation A → Ablation B), supporting SatCLIP as an informative semantic prior under 2023 shift.

XGBoost comparability. XGBoost is trained/evaluated on a simpler target (the mean of the 24-hour horizon), so its lower MAE is not directly comparable to full-trajectory forecasting.

Spatial holdout (cold start, unseen stations). Table 2 reports performance on held-out stations under a 20% station holdout split.

Table 2. 2023 Performance under Station Holdout (20% Unseen Stations)

Model	MAE	RMSE	R^2
Graph WaveNet (PM-only)	3.6786	8.2124	0.3765
SatCLIP-LSTM-GNN (Ours)	**3.5371**	**8.0594**	**0.3995**

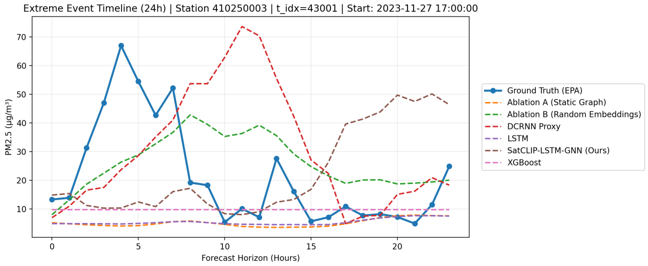

Fig. 2. Extreme-event case study (24-hour horizon). Ground-truth $\text{PM}_{2.5}$ at Station 410250003 for the critical episode starting at time index t = 43001 (Start: 2023-11-27 17:00), compared against forecasts from selected model variants.

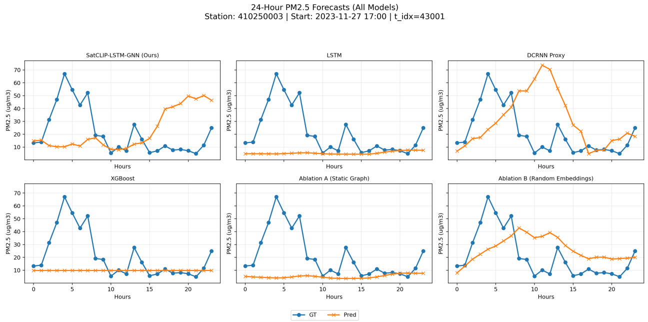

Fig. 3. Selected-model comparison for the critical episode (Time Index t = 43001, Station 410250003). In this specific window, Ablation B (Random Embeddings) tracks the early rise/relaxation more closely than the aggregate winner (Ablation A), illustrating that episode-level fidelity can diverge from average test performance.

Cold-start advantage. Although Graph WaveNet is slightly stronger on the standard 2023 seen-station evaluation (Table 1), SatCLIP improves generalization to unseen stations: MAE decreases from 3.6786 to 3.5371 on held-out stations (∼3.8% improvement), supporting frozen geospatial foundation priors in deployment settings where new sensors are added without retraining.

Extreme-event case study (single episode). Aggregate metrics

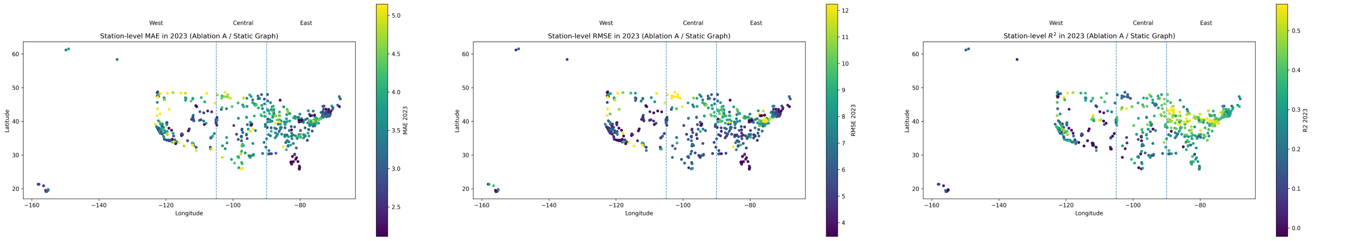
West
Central
East
Station-level MAE in 2023 (Ablation A / Static Graph)
Latitude
Longitude
60
50
40
30
20
−160
−140
−120
−100
−80
MAE 2023
5.0
4.5
4.0
3.5
3.0
2.5
Station-level RMSE in 2023 (Ablation A / Static Graph)
RMSE 2023
12
11
10
9
8
7
6
5
4
Station-level R^2 in 2023 (Ablation A / Static Graph)
R2 2023
0.5
0.4
0.3
0.2
0.1
0.0

Fig. 4. **Geographic error landscape in 2023 (Ablation A / Static Graph).** Station-level MAE, RMSE, and R^2 (left→right). Dashed lines denote coarse West/Central/East longitude bins.

can conceal behavior on rare high-impact episodes; we therefore include an illustrative analysis at Station 410250003 and time index t=43001 (Figs. 2–3). In this window, variant ranking can differ from aggregate performance, motivating future stratified evaluation over multiple extreme episodes.

Geographic performance. Fig. 4 summarizes station-level MAE/RMSE/R^2 for Ablation A in 2023. Errors are spatially heterogeneous, with a tendency toward higher error and lower R^2 in parts of the West/Central relative to the East (summarized in Fig. 5), motivating region-aware reporting and transfer tests.

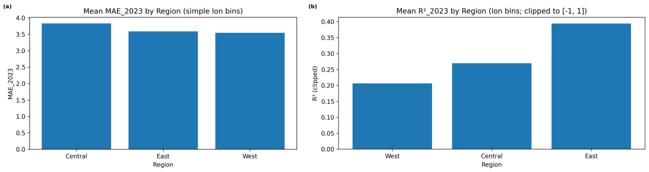

Fig. 5. **Regional performance in 2023 (Ablation A).** (a) Mean station-level MAE by coarse longitude-based region (West/ Central/ East). (b) Mean station-level R^2 by the same regions (clipped to $[-1, 1]$).

Static vs. wind gating. Despite being physics-inspired, wind gating underperforms stable static connectivity on aggregate 2023 metrics (Table 1). A plausible explanation is brittleness from hard thresholding and wind-field mismatch (coarse/imputed winds), plus temporal misalignment between instantaneous winds and multi-hour transport; these motivate future soft/learned gating alternatives. Fig. 6 provides structural intuition: under the chosen (τ, ε), wind gating can yield a much denser and time-varying adjacency than the static radius graph, increasing sensitivity to wind noise, thresholding, and temporal mismatch.

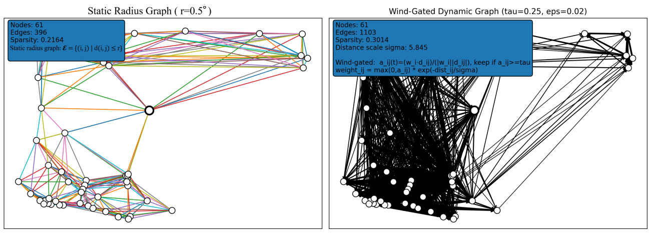
Static Radius Graph (r=0.5°)
Nodes: 61
Edges: 396
Sparsity: 0.2164
Static radius graph: $\mathcal{E} = \{(i,j) \mid d(i,j) \leq r\}$
Wind-Gated Dynamic Graph (tau=0.25, eps=0.02)
Nodes: 61
Edges: 1103
Sparsity: 0.3014
Distance scale sigma: 5.845
Wind-gated: a_ij(t)=(w_i·d_ij)/(|w_i||d_ij|), keep if a_ij>=tau
weight_ij = max(0,a_ij) * exp(-dist_ij/sigma)

Fig. 6. Local neighborhood graphs for Station 410250003 at t=43001. Left: static radius graph (r=0.5°). Right: wind-gated dynamic graph with distance-decayed weights (τ=0.25, ε=0.02). Wind gating yields a denser, time-varying adjacency and can be more sensitive to wind-field noise.

8. DISCUSSION

Seen vs. unseen stations. Table 1 shows that adaptive baselines (Graph WaveNet) are strong when test stations are drawn from the same fixed network, likely because learned node embeddings and adaptive adjacency can specialize to the training layout. In contrast, Table 2 isolates the *deployment* regime where stations are excluded from training: here, SatCLIP improves MAE on unseen stations, consistent with the view that satellite-derived location semantics provide transferable context when learned station identifiers are unavailable.

Why wind gating can hurt. Although wind gating is physically motivated, hard thresholding based on instantaneous winds can introduce brittleness under noisy or coarsely resolved meteorology and can change connectivity sharply over time. This is consistent with our observation that wind-gated connectivity does not reliably improve 2023 generalization relative to the stable static radius graph (Table 1) despite potentially lower training loss in some runs. Practically simple, stable spatial priors are preferable unless wind signals are high-quality and the gating mechanism is robust.

Interpreting correlation-based metrics. Across all models, R^2 values remain moderate (Table 1). This is partly expected: $PM_{2.5}$ distributions are heavy-tailed and episodic, so correlation-based indices such as R^2 and Pearson CC are disproportionately influenced by the large low-concentration majority and can underrepresent a model's ability to track high-impact excursions. Error-magnitude metrics (MAE, RMSE) more directly reflect forecast utility for public-health applications. The moderate R^2 does not invalidate the models but reflects a recognized limitation of correlation indices when applied to skewed environmental variables.

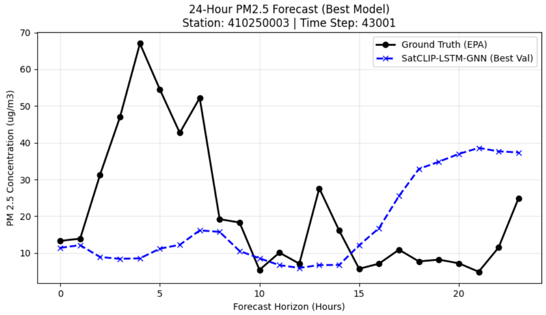

Fig. 7. Best-checkpoint forecast: dynamic trajectory with smoothing and timing errors on extreme peaks.

9. DIAGNOSTIC VISUALIZATIONS

To contextualize model behavior, we include two qualitative diagnostics. **Mean-reversion failure:** Fig. 9 shows a collapse toward a near-constant trajectory, consistent with conservative mean reversion. **Best-checkpoint response:** Fig. 7 shows a dynamic forecast (still with smoothing and peak timing/amplitude errors), illustrating model can produce non-trivial trajectories despite extreme events.

10. LIMITATIONS AND PLANNED REVISIONS

This paper shows that frozen SatCLIP priors can improve cold-start generalization to unseen stations and that stable static graphs are competitive; however, several gaps remain.

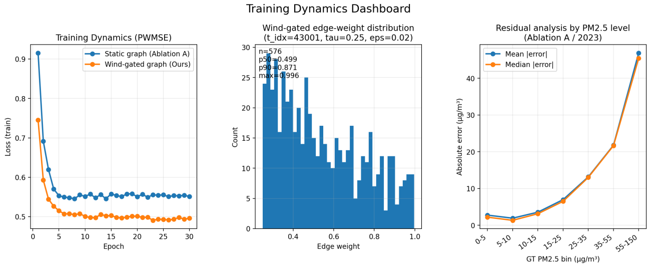

Fig. 8. Diagnostics for static vs. wind-gated connectivity: training loss (PWMSE), wind-gated edge-weight distribution at $t_{\mathrm{idx}} = 43001$ after thresholding ($\tau = 0.25$, $\epsilon = 0.02$), and residual error stratified by ground-truth $\mathrm{PM}_{2.5}$ bins.

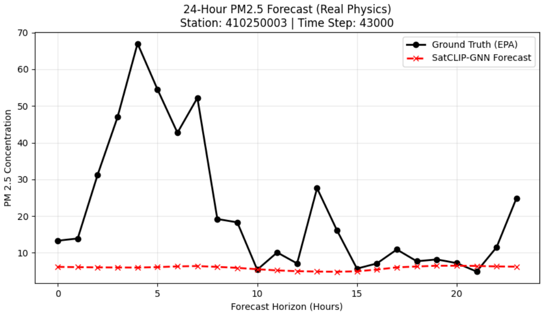

Fig. 9. Example failure mode: near-flat mean-reverting forecast over a 24-hour horizon.

Wind gating. The wind-gated graph underperformed static connectivity in 2023; likely contributors include wind-field noise, hard-thresholding brittleness, and temporal misalignment. *Planned:* replace hard gating with soft/continuous weights; evaluate seasonal subsets; and quantify robustness to missing or noisy winds.

Wind altitude. The current pipeline uses ERA5-Land 10 m winds, which are strongly influenced by local surface roughness, terrain channeling, and sub-grid obstructions, making spatial interpolation to station locations less reliable. Pressure-level winds (e.g., 925 hPa or 850 hPa from ERA5 upper-air fields) better represent synoptic-scale transport and are less sensitive to surface heterogeneity; incorporating these is planned for future investigation.

Baselines and attribution. While we include Graph WaveNet under the same 24→24 protocol (PM-only and optional PM+Met; Table 1), additional modern ST-GNN baselines (e.g., ASTGCN, AGCRN) under identical preprocessing/normalization would further strengthen the study.

Generalization beyond station holdout. We report a 20% station holdout split (Table 2), but broader transfer remains open. *Planned:* geographic holdouts (region/state), stronger distribution shifts (e.g., pre-smoke → smoke-heavy periods), and multi-seed holdout splits.

Extreme-event quantification. The current episode analysis is illustrative (single event; Section 7). *Planned:* evaluate 5–10 representative extreme events, report stratified metrics (e.g., MAE/RMSE above the 90th percentile), and add calibration-style diagnostics for peak under/overprediction.

Statistical validation and sensitivity. We currently report single-run metrics. *Planned:* multiple seeds with confidence intervals, sensitivity over radius r and gating thresholds, and paired bootstrap/permutation testing over station-day samples.

Reproducibility. We report the exact chronological split (2019–2021/2022/2023), the station-holdout protocol (20% unseen stations), the fixed input/output horizons (24→24), and normalization rules (training-only statistics). These choices are intended to reduce evaluation leakage and make results comparable across baselines and variants.

11. CONCLUSION

We studied $\mathrm{PM}_{2.5}$ forecasting with foundation-model location priors and graph learning under a realistic chronological evaluation. On the standard 2023 test set over seen stations, Graph WaveNet provides slightly stronger aggregate performance under a minimal $\mathrm{PM}_{2.5}$-only "proper baseline" setting, while our SatCLIP + static radius-graph variant remains competitive and improves MAE by 32.0% relative to a controlled DCRNN-style proxy baseline. Crucially, under a station holdout protocol that evaluates *unseen sensors*, SatCLIP improves MAE on held-out stations from 3.679 (Graph WaveNet) to 3.537, demonstrating a practical cold-start advantage for new deployments. Training curves further show that wind-gated connectivity can reach lower training PWMSE without improving 2023 generalization, suggesting that hard physics-inspired masking may be brittle under imperfect wind fields. Future work includes geographic transfer tests, extreme-event quantification, and stronger ST-GNN baselines with statistical validation.

12. ACKNOWLEDGEMENT

The research is supported in part by NSERC Collaborative Research and Training Experience (CREATE 596346-2025).

13. REFERENCES

[1] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

[2] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang, "Graph wavenet for deep spatial-temporal graph modeling," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[3] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," *arXiv preprint arXiv:2005.11650*, 2020.

[4] Bing Yu, Haoteng Yin, and Zhanxing Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.

[5] Shengnan Guo, Youfang Lin, Feng, et al., "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[6] Lian Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Alec Radford, Kim, et al., "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[8] K. M. Klemmer, E. Kuhn, and S. Zhu, "SatCLIP: Global, general-purpose location embeddings with satellite imagery," *arXiv preprint arXiv:2311.17179*, 2023.