AIRQUALITYBENCH: A REALISTIC EVALUATION BENCHMARK FOR GLOBAL AIR QUALITY FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Air-quality forecasting models are commonly evaluated on regional, preprocessed, and normalized datasets, where missing observations are removed or artificially completed. Such protocols simplify comparison but hide the conditions that dominate real monitoring networks: uneven global coverage, structured missingness, heterogeneous pollutant scales, and deployment cost. We introduce **AirQualityBench**, a global multi-pollutant benchmark designed to evaluate forecasting models under these realistic conditions. The benchmark contains hourly observations from 3,720 monitoring stations over 2021–2025, covers six major pollutants, and preserves provider-native observation masks. Rather than imputing a dense data tensor, AirQualityBench exposes missingness as part of the forecasting problem and reports errors on valid future observations after inverse transformation to physical concentration scales. Evaluating representative spatio-temporal models under this unified protocol shows that strong performance on sanitized datasets does not reliably transfer to global, fragmented monitoring streams. AirQualityBench therefore serves as a realistic testbed for scalable, mask-aware, and physically interpretable air-quality forecasting. All benchmark data, code, evaluation scripts, and baseline implementations are available at https://anonymous.4open.science/r/AirQualityBench-85B1/.

1 INTRODUCTION

Air-quality forecasting is a challenging spatio-temporal prediction problem because pollutant dynamics are shaped by regional transport, meteorology, emissions, and heterogeneous monitoring coverage (Council et al., 2010; Du et al., 2019; Organization et al., 2021; Chen et al., 2025). Recent deep spatio-temporal models have achieved strong results on curated regional datasets, but their progress depends critically on whether the underlying benchmarks reflect real monitoring conditions. The rapid evolution of these predictive architectures relies heavily on rigorous benchmarks, which serve as the "yardstick" for measuring algorithmic progress and ensuring reproducibility (Liu et al., 2023). High-quality air quality benchmarks are particularly vital as they provide the standardized environment necessary to validate a model's ability to generalize across diverse urban topologies and varying climatic conditions. In the field of spatio-temporal learning, established datasets have historically catalyzed breakthroughs by providing a common ground for comparing disparate neural architectures, such as Graph Convolutional Networks (GCNs) and Transformers, thereby bridging the gap between theoretical research and real-world deployment.

Despite their value, existing public air-quality forecasting benchmarks often simplify several aspects that are central to deployment-oriented evaluation. **First**, existing benchmarks (Wang et al., 2020; Liu et al., 2021; Wang et al., 2025) are geographically confined to single cities or specific administrative regions with limited node scales, which effectively severs the planetary-scale transport patterns of pollutants and prevents models from validating their scalability in large-scale networks. **Second**, these benchmarks frequently employ manual interventions to construct "sanitized" data tensors, such as linear interpolation or mean imputation. This deviates from the physical reality of *Missing Not At Random* (MNAR) patterns (Vicent et al., 2026), leading to a pervasive synthetic bias that fails to reflect model robustness against fragmented, real-world observation streams. **Third**, the prevailing reliance on unitless, normalized evaluation metrics obscures the physical interpretability

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

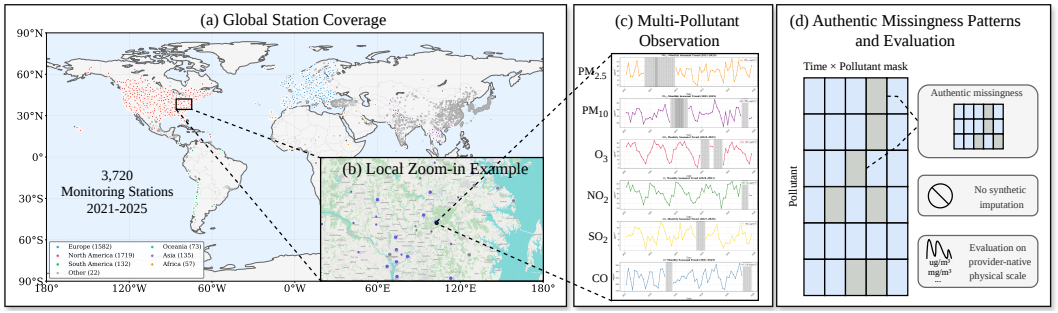

Figure 1: **Overview of AirQualityBench.** The benchmark combines a global network of 3,720 monitoring stations, synchronized observations of six pollutants, authentic missingness patterns, and physical-scale evaluation, providing a realistic testbed for large-scale spatio-temporal air quality forecasting.

and toxicological significance of prediction errors (Li et al., 2023; Hunt, 2025), creating a profound gap between laboratory experiments and the requirements of environmental regulatory deployment.

We argue that the central obstacle is not only the lack of a larger air-quality dataset, but the lack of a realistic evaluation protocol. Existing benchmarks often simplify the monitoring process into dense or partially completed tensors and report errors on normalized scales. While useful for controlled experiments, this setting removes exactly the factors that determine whether a forecasting model can be deployed in real sensing networks: missing observations, non-uniform spatial coverage, pollutant-specific scales, and computational cost. AirQualityBench is designed around these factors. It evaluates whether models can forecast from fragmented multi-pollutant histories, remain robust across uneven global monitoring networks, and produce errors that are interpretable in physical concentration units.

To support realistic and reproducible evaluation of air-quality forecasting models, we introduce **AirQualityBench**, a global multi-pollutant benchmark with authentic missingness and physical-scale reporting (Fig. 1). Our contributions are threefold:

- **Global multi-pollutant benchmark.** AirQualityBench contains hourly observations from 3,720 resident monitoring stations over 2021–2025, covering $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, and CO. It provides a geographically diverse testbed for evaluating scalability and cross-region heterogeneity beyond regional benchmarks.
- **Mask-aware physical-scale evaluation.** We preserve provider-native observation masks instead of constructing fully imputed tensors, and evaluate predictions only on valid future measurements. Errors are reported after inverse transformation to pollutant-specific concentration scales, enabling physically interpretable metric.
- **Unified baselines and diagnostics.** We evaluate representative spatio-temporal forecasting models under the same split, masking protocol, graph construction, and scripts, revealing their behavior under pollutant-specific sparsity, non-uniform station coverage, and accuracy–efficiency trade-offs.

2 Problem Definition and Related Work

In this section, we formally define the masked spatio-temporal forecasting task and review the relevant literature in deep spatio-temporal learning and air quality modeling.

2.1 Problem Definition

Task Formalization. We represent the global monitoring network as a spatio-temporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ denotes the set of $N = 3,720$ stations and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix. At each time step t, the concentrations of $C = 6$ pollutants and their corresponding availability

are represented by an observation tensor $\mathbf{X}_t \in \mathbb{R}^{N\times C}$ and a binary mask tensor $\mathbf{M}_t \in \{0,1\}^{N\times C}$. Given a historical window T, the goal is to learn a mapping f_θ with the following input-output configuration:

$$f_\theta : \underbrace{\mathbb{R}^{T\times N\times C} \times \{0,1\}^{T\times N\times C}}_{\text{Historical Observations \& Masks}} \longrightarrow \underbrace{\mathbb{R}^{T'\times N\times C}}_{\text{Future Predictions}} \tag{1}$$

where T' is the forecasting horizon. This formulation forces the model to reason directly from fragmented historical streams to generate complete future trajectories.

Physical-Scale Evaluation. AirQualityBench evaluates predictions after inverse transformation to the provider-native concentration scale. We report masked MAE, MSE, RMSE, and MAPE on valid future observations. Since the benchmark contains heterogeneous pollutants and provider-specific unit conventions, pollutant-wise metrics are the primary physically interpretable results. Cross-pollutant global aggregates are provided only as benchmark-level ranking summaries and should not be interpreted as a single physically homogeneous concentration error.

2.2 Related Work

Deep spatio-temporal forecasting. Deep spatio-temporal forecasting has been widely studied in networked sensing systems, including traffic, mobility, and environmental monitoring. Representative models range from graph-recurrent and graph-convolutional methods, such as DCRNN (Li et al., 2017), STGCN (Yu et al., 2017), and GWN (Wu et al., 2019), to attention-based and Transformer-style architectures, such as ASTGCN (Guo et al., 2019), STTN (Xu et al., 2020), and PDFormer (Jiang et al., 2023). Recent methods further explore adaptive graph learning, decoupled spatial–temporal dynamics, scalable temporal modeling, and event-aware forecasting, including AGCRN (Bai et al., 2020), D^{2}STGNN (Shao et al., 2022), MAGE (Ma et al., 2025a), BiST (Ma et al., 2025b), and IGSTGNN (Fan et al., 2026). For air-quality forecasting, models such as AirFormer (Liang et al., 2023) and PCDCNet (Wang et al., 2025) further incorporate meteorological information, geographical priors, or physical–chemical constraints. Despite these advances, most evaluations remain tied to regional, partially completed, or normalized datasets, leaving model robustness under global-scale, pollutant-specific, and authentically fragmented monitoring streams insufficiently examined.

Air-quality benchmarks and open observation platforms. Public datasets and benchmarks have played an important role in air-quality forecasting research. KDD Cup 2018 (Liu et al., 2021) provides a widely used regional benchmark, while KnowAir (Wang et al., 2020) and KnowAir-V2 (Wang et al., 2025) offer curated resources with meteorological or physical–chemical information. OpenAQ (Hasenkopf et al., 2015) aggregates global air-quality observations from heterogeneous monitoring providers and serves as a key open data platform. However, existing forecasting benchmarks are still commonly limited to specific cities, countries, or administrative regions, focus on a small set of pollutants, rely on interpolation or imputation to construct clean tensors, or report errors mainly in normalized spaces. These design choices facilitate controlled comparison but underrepresent the fragmented, heterogeneous, and physically grounded conditions of operational monitoring networks.

3 Limitations of Existing Air-Quality Forecasting Benchmarks

Existing air-quality forecasting benchmarks are limited not only by dataset size, but also by the evaluation assumptions they encode. As summarized in Table 1, many prior benchmarks are regional in spatial scope, partially sanitized through interpolation or completion, and evaluated primarily on normalized scales. These choices simplify controlled experimentation, but they also move the benchmark setting away from the fragmented, heterogeneous, and physically grounded conditions encountered in real monitoring networks. Below, we highlight three major limitations that motivate the design of AirQualityBench.

Limited spatial scope. A first limitation of existing benchmarks is their restricted geographic coverage. Most widely used datasets are constructed within a single city, country, or administrative region, with limited station counts and spatial diversity. While such datasets are valuable for studying regional forecasting, they provide limited evidence about whether a model can scale to large, non-uniform monitoring graphs or remain robust across regions with different sensing density, climate

Table 1: Comparison of standard air quality spatio-temporal benchmarks.

Dataset	Spatial Scale	Nodes (N)	Time Span	Pollutants
GLB-$PM_{2.5}$ (Liu et al., 2021)	Regional (CN)	534	2018/09–2018/12	$PM_{2.5}$
KnowAir (Wang et al., 2020)	Regional (CN)	184	2015/01–2018/12	$PM_{2.5}$, O_3
KnowAir-V2 (Wang et al., 2025)	Regional (CN)	355	2016/01–2023/12	$PM_{2.5}$, O_3
AirQualityBench	**Global**	**3,720**	**2021/01–2025/12**	**All 6 pollutants**

Note: "All 6 pollutants" denotes $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, and CO. Particulate matter is predominantly reported in $\mu g/m^3$; gaseous pollutants follow the provider-native parameter.units metadata retained by preprocessing.

conditions, emission sources, and pollutant dynamics. Strong performance on region-bounded datasets therefore does not necessarily imply robustness under broader deployment settings.

Sanitized missingness. A second limitation lies in how many benchmarks handle incomplete observations. A common practice is to construct relatively clean data tensors through interpolation, imputation, or other forms of preprocessing before evaluation. Although convenient, this removes an essential characteristic of real monitoring systems, where observations are often fragmented, sparse, and missing in structured rather than purely random ways. Consequently, models evaluated on completed tensors may perform well in laboratory-style settings while remaining fragile when faced with authentic sensing streams.

Weak physical interpretability. A third limitation is the widespread reliance on normalized or standardized evaluation spaces. While normalization is convenient for optimization and comparison, it obscures the practical meaning of prediction error and weakens the connection between benchmark performance and real environmental decision-making. In operational settings, forecasting error is ultimately understood in provider-native concentration units rather than dimensionless normalized residuals. A realistic benchmark should therefore emphasize physical-scale evaluation so that reported errors remain directly interpretable in terms of actual pollutant concentrations (Hunt, 2025).

Implications and new modeling challenges. Taken together, these limitations define a broader gap between existing benchmarks and realistic air-quality forecasting conditions. AirQualityBench is designed to close this gap through global-scale coverage, preservation of authentic observation incompleteness, multi-pollutant forecasting, and evaluation on the original physical scale. These choices do not merely enlarge the dataset, they also introduce new modeling challenges. First, models must learn spatial dependencies across a globally distributed and highly non-uniform monitoring network, where useful signals may arise from both nearby stations and longer-range transport patterns. Second, they must forecast from structured missingness rather than from completed tensors, requiring robustness to pollutant-specific sparsity and station-level observation gaps. Third, they must handle heterogeneous pollutant dynamics and provider-native concentration scales, making pollutant-wise physical-scale accuracy as important as aggregate benchmark ranking. In this sense, AirQualityBench is intended not only as a larger dataset, but also as a stress test for whether spatio-temporal models remain reliable under deployment-relevant sensing conditions.

4 AirQualityBench Dataset

In this section, we detail the rigorous engineering pipeline used to construct **AirQualityBench**, moving from raw global observation streams to a standardized, multi-pollutant spatio-temporal benchmark.

4.1 Data Acquisition and Organization

This subsection summarizes how we convert raw global observations into a benchmark-ready spatio-temporal dataset. We focus on three components: station selection and organization, physically faithful evaluation under authentic missingness, and a standardized geographic graph prior.

Dataset acquisition and filter. The raw data for AirQualityBench is harvested from the **OpenAQ** (Hasenkopf et al., 2015) platform, which aggregates hourly pollutant concentrations from

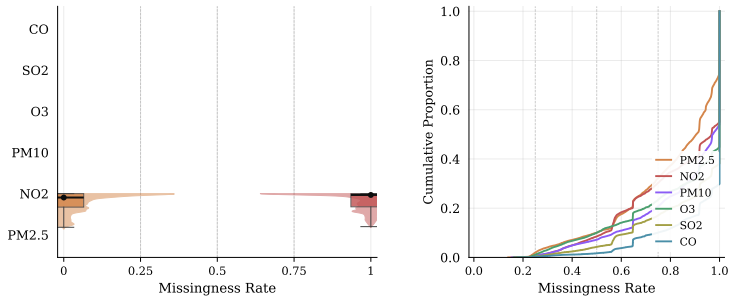

Figure 2: **Pollutant-specific missingness in AirQualityBench.** Station-level missingness distributions show heterogeneous coverage across pollutants, with substantially sparser observations for gaseous species.

diverse global monitoring networks. We retain data from **January 1, 2021, to December 31, 2025** and apply a resident station filter: a station is kept if it maintains a valid observation rate above **50%** for at least one pollutant over the full five-year span. This criterion yields **3,720 resident stations**. Coverage remains heterogeneous across pollutants, with $PM_{2.5}$ having the broadest active station set ($n = 3,412$) and CO the sparsest ($n = 1,456$). We use $N = 3,720$ for the total resident station count and n for pollutant-specific active stations. The 50% threshold is a pragmatic compromise between spatial coverage and observational reliability, retaining about 35% of candidate stations while discarding highly intermittent sites.

Physical-scale evaluation protocol. Each sensor record retains provider-reported `parameter.units` metadata, and we preserve this provider-native physical scale throughout preprocessing. After temporal alignment to UTC and basic quality control, we remove sentinel values, negative readings, and physically implausible outliers, but we do not replace missing values with interpolation. Instead, we use an authentic masking strategy with a binary mask tensor $\mathbf{M} \in \{0,1\}^{T\times N\times C}$, where $M_{t,n,c} = 1$ denotes a valid observation. AirQualityBench therefore mandates evaluation on the original physical scale: models may normalize inputs internally, but all reported metrics are inverse-transformed back to the provider-native units of $PM_{2.5}, PM_{10}, O_3, NO_2, SO_2$, and CO.

Spherical graph topology. To provide a standardized spatial prior for all graph-based baselines, we construct a spherical k-NN graph from station coordinates using Haversine distance. This design accounts for the Earth's curvature and avoids treating the global monitoring network as a planar grid. Each station is connected to its k nearest geographic neighbors, yielding a reproducible topology for evaluating spatial dependency modeling across continents. The detailed distance formula and adjacency construction are provided in Appendix B.

4.2 Dataset Analysis

AirQualityBench is not only large in scale but also structurally diverse. Its descriptive statistics reveal three properties that are central to benchmark difficulty: heterogeneous sensing coverage across pollutants, geographically meaningful spatial dependence, and persistent multi-scale temporal dynamics. Together, these properties explain why the dataset is challenging while still remaining learnable.

Structured missingness is a defining property of AirQualityBench. As shown in Figure 2, the six pollutants exhibit markedly different missingness profiles. $PM_{2.5}$ and PM_{10} display broad, right-skewed distributions spanning the full [0, 1] interval, indicating heterogeneous but non-negligible coverage across stations. Gaseous species, by contrast, are far more fragmented: O_3, SO_2, and CO concentrate near 1.0, while NO_2 reveals a bimodal structure with a subset of stations achieving substantially better coverage. This pollutant-specific sparsity structure—rather than a single homogeneous masking assumption—is exactly why we preserve authentic missingness and report pollutant-wise results: robustness under fragmented sensing is a central challenge, not a nuisance to be hidden by interpolation.

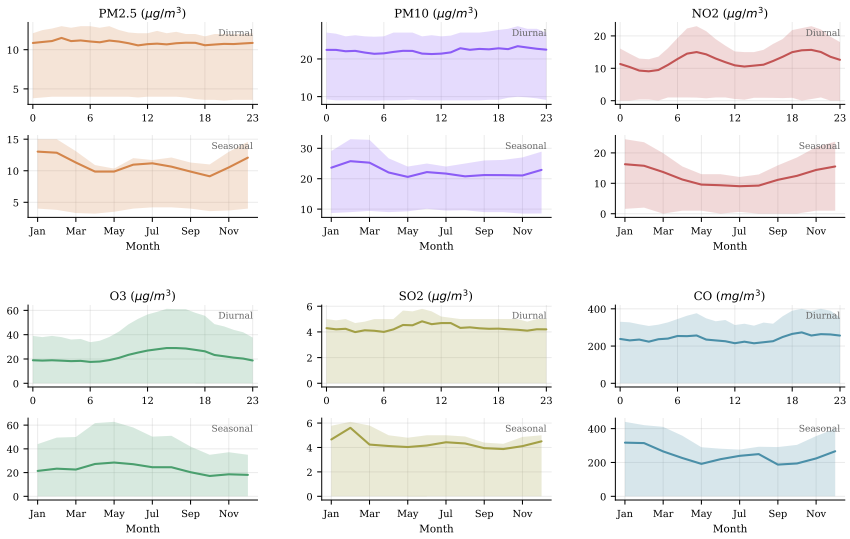

Figure 3: **Multi-scale temporal dynamics in AirQualityBench.** Diurnal and seasonal climatologies show that all six pollutants preserve regular temporal structure with pollutant-dependent rhythms.

Spatial and temporal structure under global fragmentation. Despite the planetary-scale coverage of AirQualityBench, pairwise correlations exhibit a systematic decay with Haversine distance rather than collapsing into spatially unstructured noise (Appendix B). This pattern supports the use of a standardized spherical k-NN graph as a principled inductive bias. Meanwhile, the decay is gradual rather than sharply local, suggesting that informative dependencies can extend beyond immediate neighbors and making long-range spatio-temporal forecasting particularly challenging. AirQualityBench also preserves clear temporal structure: as shown in Figure 3, all pollutants exhibit diurnal and seasonal regularities, although the dominant rhythms differ across pollutants. These observations indicate that the benchmark is not merely sparse, but contains learnable spatial and temporal signals under authentic fragmentation. For traceability, we retain provider-native unit metadata; the handling of CO unit heterogeneity is detailed in Appendix F.

Data and code availability. We will release the processed benchmark tensors, authentic observation masks, station metadata, spherical k-NN graph, chronological splits, evaluation scripts, and baseline implementations. The release is designed to make all reported benchmark results reproducible from the same preprocessing and evaluation pipeline. The raw observations are derived from OpenAQ, and users should follow the original provider terms when accessing raw records.

5 Benchmark Protocol and Results

5.1 Experimental Setup

Datasets and protocol. We evaluate all models on **AirQualityBench** under a chronological split designed to test generalization across unseen climatic cycles. Specifically, data from **2021–2023** are used for training, **2024** for validation, and **2025** for testing, so that the test set covers a full unseen annual cycle of seasonal transitions, extreme weather events, and pollution episodes. The benchmark contains hourly observation streams from **3,720 resident stations** worldwide, covering six pollutants: $PM_{2.5}, PM_{10}, O_3, NO_2, SO_2$, and CO. Throughout all experiments, we strictly follow the **Authentic Masking** protocol: models only observe physically valid measurements, while missing values and outliers are explicitly masked during both training and evaluation.

Baselines. We compare against representative spatio-temporal forecasting baselines spanning multiple architectural families, including graph-based methods (**DCRNN** (Li et al., 2017), **AGCRN** (Bai et al., 2020)), temporal convolution methods (**STGCN** (Yu et al., 2017), **GWN** (Wu et al., 2019)), attention-based methods (**ASTGCN** (Guo et al., 2019), **STTN** (Xu et al., 2020), **PDFormer** (Jiang et al., 2023)), and efficiency-oriented models (**MAGE** (Ma et al., 2025a), **BiST** (Ma et al., 2025b)). We

Table 2: Main benchmark results using released global aggregate metrics across forecasting horizons. These aggregates are used for model ranking rather than single-unit physical interpretation. Lower is better. Best results are in **bold** and second-best results are underlined.

Method	Horizon 6			Horizon 12			Horizon 18			Horizon 24		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
DCRNN	6.04	47.42	0.80	7.27	56.07	0.97	7.73	58.65	1.06	7.48	52.11	0.97
STGCN	7.15	53.09	1.08	7.60	54.89	1.08	7.99	56.45	1.16	8.26	55.07	1.13
GWN	6.09	49.13	0.81	6.87	49.94	0.90	7.42	55.47	1.04	**6.98**	**50.05**	**0.81**
ASTGCN	12.69	72.87	1.40	13.73	69.55	13.43	12.99	75.38	1.66	12.82	79.36	1.61
AGCRN	11.11	69.27	1.85	12.14	80.10	1.85	12.24	79.54	2.10	7.59	53.17	1.02
STTN	5.90	**41.41**	1.27	**6.56**	**46.54**	1.45	**6.83**	**48.58**	1.24	7.22	53.52	1.44
DSTAGNN	8.81	54.66	1.16	9.40	57.63	1.32	10.63	65.76	1.45	11.03	70.57	1.87
D2STGNN	**5.88**	43.29	**0.71**	6.62	49.11	0.90	6.99	51.20	**0.85**	7.27	53.70	1.01
PDFormer	7.20	50.99	0.81	7.64	52.93	**0.82**	8.05	58.32	0.92	8.09	58.57	0.94
MAGE	6.46	49.15	1.29	7.22	53.10	1.79	7.78	53.61	1.75	7.83	56.78	1.81
BiST	12.75	86.32	4.76	17.19	94.41	9.06	19.32	99.92	8.87	19.28	100.80	8.35
IGSTGNN	18.90	97.73	4.11	23.28	101.85	13.65	23.29	102.72	13.46	23.29	103.74	14.06

All entries are global aggregate metrics from the benchmark evaluation release. Since global averages denormalized errors across heterogeneous pollutants and unit conventions, it should be interpreted as a benchmark ranking statistic rather than as a physically homogeneous MAE/RMSE. Pollutant-wise physically interpretable results are provided in Appendix D.

further include **D2STGNN** (Shao et al., 2022) and **IGSTGNN** (Fan et al., 2026) to cover decoupled spatial–temporal modeling and event-aware forecasting paradigms.

Baseline adaptation and fairness. All baselines are evaluated using the same chronological split, forecasting horizons, station set, pollutant set, spherical k-NN graph, and authentic evaluation mask. For models that do not natively support missing inputs, missing values are filled with zeros after normalization and the corresponding binary masks are provided as additional input indicators when applicable. Training losses and evaluation metrics are computed only on valid observations. This protocol ensures that performance differences mainly reflect model behavior under the same fragmented input streams rather than differences in preprocessing or data completion.

Evaluation metrics. We follow a strict **physical-scale evaluation** protocol: predictions are first inverse-transformed to provider-native units, and metrics are computed only on valid observations indicated by the authentic mask. We report four standard forecasting metrics: MAE, MSE, RMSE, and MAPE. In the main paper, Table 2 summarizes horizon-wise global performance using MAE, RMSE, and MAPE. Because the global field aggregates denormalized errors across heterogeneous pollutants, these scores are intended as benchmark-level ranking summaries rather than single-unit physical quantities. Pollutant-wise results and complementary statistics are provided in Appendix D and Appendix G.

5.2 Performance Comparison

Unless otherwise stated, all numbers in this section correspond to metrics computed on denormalized predictions and labels under the released evaluation protocol. Table 2 reports the main benchmark table. We keep the primary comparison focused on overall ranking and place the three global metrics in a single horizon-grouped table.

Beyond the absolute rankings, the results reveal how AirQualityBench stresses different spatio-temporal inductive biases rather than simply rewarding architectural complexity. Models with explicit spatio-temporal dependency modeling and decoupled spatial–temporal dynamics form the strongest overall tier, suggesting that robust interaction modeling is important when forecasting from fragmented multi-pollutant observations. Temporal-convolution-based methods are also highly competitive, especially at longer horizons, indicating that the benchmark preserves strong diurnal and seasonal regularities that can be exploited by stable temporal backbones. In contrast, more flexible adaptive or attention-heavy designs do not always improve performance, which suggests

Table 3: Computational efficiency profiling on AirQualityBench ($N = 3,720$). Lower inference time and memory are better.

Method	Params (M)	Infer (ms)	Memory (MB)
STGCN	1.68	3.60	412.18
DCRNN	0.04	21.15	722.17
GWN	0.39	16.33	1162.29
AGCRN	0.11	15.82	6594.34
PDFormer	0.15	5.54	1108.13
ASTGCN	55.69	14.21	2206.23
STTN	1.31	35.37	819.24
DSTAGNN	41.57	5.61	970.80
D2STGNN	0.43	70.17	16560.93
MAGE	11.86	3.11	246.90
BiST	0.38	0.60	31.52
IGSTGNN	0.42	4.80	1055.80

"Infer" and "Memory" are measured under the same profiling pipeline for all benchmark implementations and correspond to the instantiated configurations used in our experiments. Additional ablations and seed-robustness results are moved to Appendix C.

that authentic missingness, pollutant-specific sparsity, and the sparse non-uniform global graph can make learned dependencies difficult to estimate reliably. The metric-sensitive ranking further shows that heterogeneous physical scales expose different model behaviors under MAE, RMSE, and MAPE. Overall, these results suggest that AirQualityBench differs from sanitized regional benchmarks not only in scale, but also in diagnostic value. The benchmark exposes model sensitivity to authentic missingness, pollutant-specific sparsity, heterogeneous physical scales, and the cost of scaling spatio-temporal dependency modeling to a non-uniform global monitoring network.

5.3 Computational Efficiency

Computational efficiency is a key concern for AirQualityBench because forecasting must be performed over a global network with 3,720 stations, six pollutants, and authentic missingness patterns. Unlike small regional benchmarks, such a setting can amplify the cost of recurrent message passing, adaptive graph learning, and dense spatio-temporal attention. We therefore profile all implemented baselines under the same pipeline and report parameter count, inference latency, and memory usage in Table 3. These measurements describe the instantiated configurations used in our experiments and provide a complementary view to the forecasting accuracy reported in Table 2.

Table 3 shows that computational cost is not determined by parameter count alone. Some compact models still incur substantial memory usage or latency, suggesting that runtime behavior depends strongly on the form of spatio-temporal computation rather than only on the number of trainable weights. Conversely, some models with larger parameter counts remain relatively efficient under the tested configuration. This observation is important for benchmark interpretation: a method that is accurate but memory-intensive may be difficult to deploy on large fragmented sensing networks, while a lightweight method must still preserve enough capacity to handle missingness, pollutant heterogeneity, and long-range spatial dependencies.

5.4 Accuracy–Efficiency Trade-off

Figure 4 combines the forecasting results in Table 2 with the profiling results in Table 3. By placing each model according to global MAE and inference latency, and encoding parameter count as bubble size, the figure summarizes the practical cost of achieving stronger benchmark performance. The resulting pattern shows that higher accuracy often comes with increased computational burden, whereas very fast models tend to sacrifice forecasting quality under authentic missingness and heterogeneous pollutant dynamics.

This trade-off highlights an important distinction between AirQualityBench and many sanitized regional benchmarks. Since the benchmark preserves large-scale topology, fragmented observations,

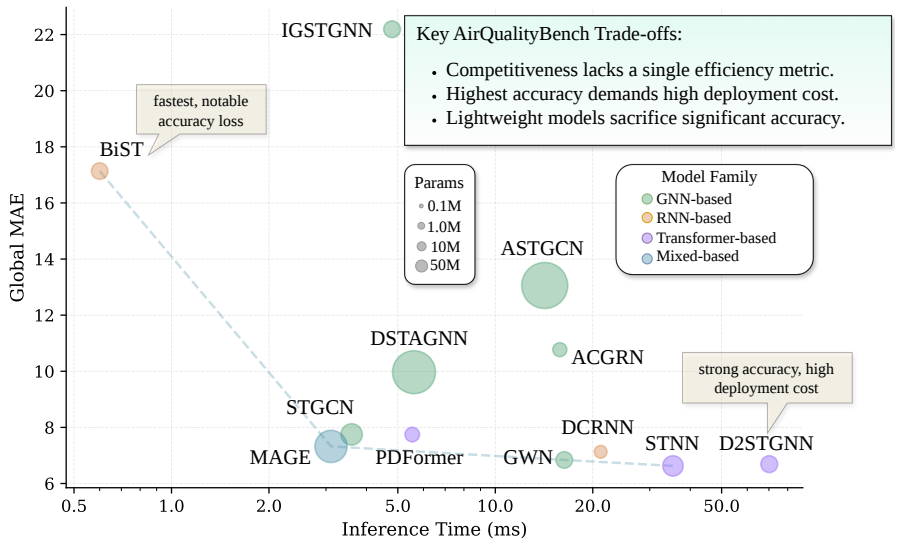

Figure 4: **Accuracy–efficiency trade-off on AirQualityBench.** Each bubble denotes a forecasting model, with position determined by global aggregate MAE and inference latency, and bubble size proportional to parameter count.

and physical-scale evaluation, practical progress cannot be assessed by accuracy alone. Instead, competitive models should move toward the favorable region of the accuracy–efficiency space: lower physical-scale error, lower inference latency, and manageable memory usage. In this sense, AirQualityBench provides a deployment-oriented testbed for studying not only whether models can forecast accurately, but also whether they can do so at a cost compatible with realistic global sensing networks.

6 Future Opportunities and Scope

AirQualityBench provides a transparent first release for global, mask-aware, physical-scale air-quality forecasting. It opens research directions in scalable graph construction, mask-aware architectures, pollutant-specific modeling, region-stratified evaluation, uncertainty-aware forecasting, and accuracy–efficiency trade-offs for large sensing networks.

Coverage imbalance. Monitoring station coverage reflects global infrastructure disparities: regions with mature monitoring systems are more densely represented, whereas under-instrumented regions have fewer stations and higher missingness. We treat spatial imbalance as a property of real monitoring infrastructure. Appendix E reports continent-level station distribution and missingness statistics, motivating region-stratified evaluation as a future benchmark track.

Provider-native unit conventions. AirQualityBench preserves provider-reported unit metadata for traceability and reproducibility, but absolute errors for gaseous pollutants, especially CO, require caution. The current release primarily supports comparative benchmarking under a fixed preprocessing pipeline. A harmonized release with station-level unit audits and explicit conversion rules remains an important extension.

Beyond supervised forecasting. The benchmark uses a fixed chronological split and authentic masking protocol. Future versions can support uncertainty estimation, intervention-aware forecasting, cross-region transfer, and event-centered evaluation, connecting spatio-temporal learning with operational air-quality decision support.

AirQualityBench is a transparent, reproducible first release rather than a final standard for all air-quality forecasting settings. By preserving authentic missingness, reporting physical-scale errors, and exposing non-uniform global coverage, it provides a foundation for developing reliable forecasting models for fragmented real-world observation streams.

AI USE STATEMENT

Generative AI tools assisted with language editing, LaTeX formatting, consistency checks, and interpretation of author-provided results. They were not used to generate datasets or experimental results. The authors are responsible for reviewing and verifying the AI-assisted content and for the final manuscript's text, claims, and artifacts.

REFERENCES

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

Chujun Chen, Weihua Chen, Linhao Guo, Yongkang Wu, Xianzhong Duan, Xuemei Wang, and Min Shao. A comprehensive review of tropospheric background ozone: definitions, estimation methods, and meta-analysis of its spatiotemporal distribution in china. *Atmospheric Chemistry and Physics*, 25(21):15145–15169, 2025.

National Research Council, Division on Earth, Life Studies, Board on Atmospheric Sciences, and Committee on the Significance of International Transport of Air Pollutants. *Global sources of local pollution: An assessment of long-range transport of key air pollutants to and from the United States*. National Academies Press, 2010.

Shengdong Du, Tianrui Li, Yan Yang, and Shi-Jinn Horng. Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 33(6): 2412–2424, 2019.

Lixiang Fan, Bohao Li, Tao Zou, Junchen Ye, and Bowen Du. Incident-guided spatiotemporal traffic forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 243–254, 2026.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 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.

Christa A Hasenkopf, JC Flasher, Olaf Veerman, and Helen Langley DeWitt. Openaq: a platform to aggregate and freely share global air quality data. In *AGU fall meeting abstracts*, volume 2015, pp. A31D–0097, 2015.

Kieran MR Hunt. Stop using root-mean-square error as a precipitation target! *arXiv preprint arXiv:2509.08369*, 2025.

Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. 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, 2023.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Zhiyuan Li, Kin-Fai Ho, Harry Fung Lee, and Steve Hung Lam Yim. Development of an integrated model framework for multi-air-pollutant exposure assessments in high-density cities and the implications for epidemiological research. *EGUsphere*, 2023:1–20, 2023.

Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 14329–14337, 2023.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems*, 36:75354–75371, 2023.

Yachuan Liu, Jiaqi Ma, Paramveer Dhillon, and Qiaozhu Mei. A new benchmark of graph learning for pm 2.5 forecasting under distribution shift. In *ACM*, pp. 6, 2021.

Jiaming Ma, Binwu Wang, Guanjun Wang, Kuo Yang, Zhengyang Zhou, Pengkun Wang, Xu Wang, and Yang Wang. Less but more: Linear adaptive graph learning empowering spatiotemporal forecasting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Jiaming Ma, Binwu Wang, Pengkun Wang, Zhengyang Zhou, Xu Wang, and Yang Wang. Bist: A lightweight and efficient bi-directional model for spatiotemporal prediction. *Proceedings of the VLDB Endowment*, 18(6):1663–1676, 2025b.

World Health Organization et al. *WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization, 2021.

Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *arXiv preprint arXiv:2206.09112*, 2022.

José F Vicent, Manuel Curado, and Marc Semper. Spatio-temporal graph neural network for inter-city air quality forecasting. *International Journal of Environmental Science and Technology*, 23(1):63, 2026.

Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting. In *Proceedings of the 28th international conference on advances in geographic information systems*, pp. 163–166, 2020.

Shuo Wang, Yun Cheng, Qingye Meng, Olga Saukh, Jiang Zhang, Jingfang Fan, Yuanting Zhang, Xingyuan Yuan, and Lothar Thiele. Pcdcnet: A surrogate model for air quality forecasting with physical-chemical dynamics and constraints. *arXiv preprint arXiv:2505.19842*, 2025.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020.

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.

APPENDIX CONTENTS

A Metric Formulations **13**

B Spherical Graph Construction and Spatial Correlation Decay **13**

C Supplementary Experiments **14**

D Per-Pollutant Detailed Results **15**

E Station Distribution Analysis **17**

F Provider-Native Units and CO Interpretation **17**

G Alternative Aggregate Metrics **18**

A METRIC FORMULATIONS

We define the masked evaluation metrics used in AirQualityBench. Let $\mathcal{O}$ denote the set of valid (observed, non-missing) entries at a given forecasting horizon, and let y_i and $\hat{y}_i$ denote the ground truth and prediction for entry $i \in \mathcal{O}$, respectively.

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} |y_i - \hat{y}_i| \tag{2}$$

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} (y_i - \hat{y}_i)^2 \tag{3}$$

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} (y_i - \hat{y}_i)^2} \tag{4}$$

Mean Absolute Percentage Error (MAPE):

$$\text{MAPE} = \frac{100}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{5}$$

Note: MAPE is undefined where $y_i = 0$. In our evaluation, we filter $y_i < \epsilon$ (with $\epsilon = 0.1$ in original units) before computing MAPE to avoid numerical instability.

B SPHERICAL GRAPH CONSTRUCTION AND SPATIAL CORRELATION DECAY

AirQualityBench provides a standardized spatial topology for graph-based forecasting models. Because the monitoring network is distributed globally, we construct the graph using spherical rather than planar distance. For two stations i and j with latitude–longitude coordinates (ϕ_i, λ_i) and (ϕ_j, λ_j), their Haversine distance is computed as

$$d_{ij} = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_i \cos\phi_j \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right), \tag{6}$$

where R is the Earth's radius, $\Delta\phi = \phi_i - \phi_j$, and $\Delta\lambda = \lambda_i - \lambda_j$. Based on these distances, we construct a binary k-nearest-neighbor graph:

$$A_{ij} = \begin{cases} 1, & \text{if } j \in \mathcal{N}_k(i), \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{N}_k(i)$ denotes the set of the k geographically closest neighbors of station i. This topology gives all graph-based baselines a reproducible spatial prior that accounts for the Earth's curvature and avoids imposing a planar-grid assumption on a planetary-scale monitoring network.

We further examine whether this distance-based prior is empirically supported by the data. Specifically, we analyze the relationship between pairwise pollutant correlation and Haversine distance between monitoring stations. Figure 5 shows that correlation generally decreases as inter-station distance increases, both in the aggregated view and across individual pollutants. This indicates that the spatial organization of AirQualityBench is not erased by its global scale: geographically closer stations remain statistically more related, while distant stations tend to exhibit weaker dependence.

This observation supports the use of the standardized spherical k-NN graph as a data-informed inductive bias rather than an arbitrary implementation choice. At the same time, the decay is gradual rather than sharply truncated: correlations do not vanish immediately outside a narrow local radius, and weak but non-negligible dependencies can persist over longer distances. Effective forecasting

models must therefore balance local spatial smoothness with the ability to capture broader-range transport and interaction effects.

We emphasize that this analysis does not assume a fixed universal decay law for all pollutants. Instead, it shows that distance-aware spatial structure remains consistently observable across the benchmark, making spherical distance-based graph construction a principled and practical choice for large-scale spatio-temporal forecasting.

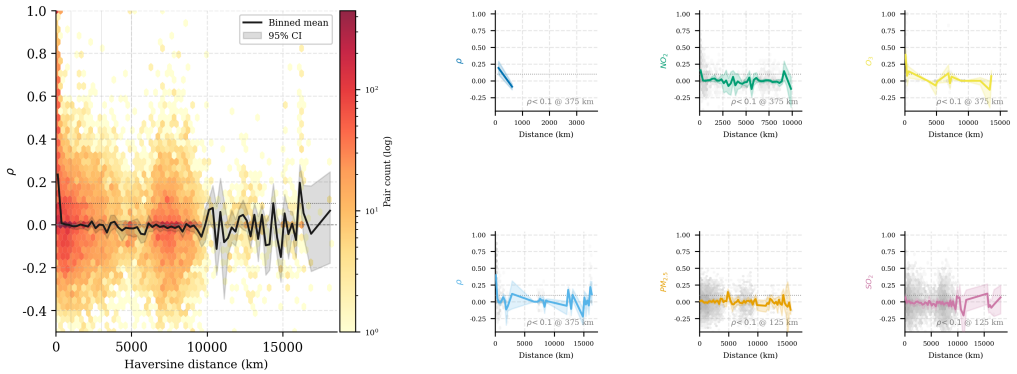

Figure 5: **Spatial correlation decay with Haversine distance.** Left: aggregated relationship between pairwise correlation and inter-station distance, summarized with distance-binned means and 95% confidence intervals. Right: pollutant-specific decay patterns for CO, NO_2, O_3, PM_{10}, $PM_{2.5}$, and SO_2. Across pollutants, pairwise correlation generally weakens with distance, supporting the use of a standardized spherical k-NN graph as a data-supported spatial inductive bias. The gradual decay further suggests that informative dependencies extend beyond strictly local neighborhoods.

C Supplementary Experiments

This section collects design ablations and seed-robustness results that extend the main paper without repeating the overall metrics already shown in Table 2.

Table 4: Ablation on k-NN graph connectivity (k) for D2STGNN at 12h horizon.

k	MAE	RMSE	MAPE
5	7.14	51.23	1.12
10	**6.62**	**49.11**	**0.90**
15	6.78	50.47	0.95
20	6.95	51.82	1.08
30	7.31	53.19	1.21

Removing the authentic mask from D2STGNN degrades MAE from 6.62 to 7.45, confirming that explicit observation-status information is useful under fragmented sensing. Replacing the fixed k-NN graph with a learned adaptive graph gives MAE 6.89, suggesting that a simple distance prior remains competitive at global scale.

Table 5: Seed robustness for D2STGNN across forecasting horizons.

Seed	6h	12h	18h	24h
42	5.88	6.62	6.99	7.27
123	5.91	6.58	7.02	7.31
456	5.85	6.65	6.95	7.24
789	5.92	6.60	7.05	7.29
1024	5.87	6.63	6.97	7.26
Mean ± Std	**5.88 ± 0.03**	**6.62 ± 0.03**	**6.99 ± 0.04**	**7.27 ± 0.03**

D PER-POLLUTANT DETAILED RESULTS

Table 6, Table 7, Table 8, and Table 9 provide pollutant-wise MAE results for all four forecasting horizons.

Table 6: Prediction performance (MAE) by pollutant type for 6h horizon.

Method	$PM_{2.5}$	PM_{10}	NO_2	O_3	SO_2	CO
D2STGNN	**2.82**	**6.02**	**3.78**	**3.83**	0.85	**43.15**
STTN	2.95	6.24	3.85	4.03	**0.81**	41.17
GWN	2.89	6.14	3.85	3.93	0.86	45.21
DCRNN	3.14	6.82	4.12	4.56	0.98	51.23
MAGE	3.08	6.72	4.08	4.78	0.95	49.87
STGCN	3.05	6.65	4.02	4.65	1.02	57.45
PDFormer	3.32	7.15	4.42	5.12	0.99	51.23
AGCRN	4.45	8.95	5.68	7.85	1.52	98.23
ASTGCN	4.78	9.45	4.25	15.67	2.28	92.45
BiST	3.95	8.45	6.12	18.45	2.85	138.92
IGSTGNN	5.87	11.25	8.75	20.12	3.45	205.67
DSTAGNN	3.52	7.45	4.58	5.34	1.21	76.23

All entries are denormalized MAE values under the benchmark evaluation protocol; pollutant units follow the original provider-native metadata retained by preprocessing.

Table 7: Prediction performance (MAE) by pollutant type for 12h horizon.

Method	$PM_{2.5}$	PM_{10}	NO_2	O_3	SO_2	CO
D2STGNN	**3.10**	**6.65**	**4.17**	**4.54**	0.95	48.48
STTN	3.13	6.75	4.22	4.72	**0.88**	**45.54**
GWN	3.20	6.87	4.30	4.64	0.96	50.84
DCRNN	3.44	7.19	4.48	4.84	1.04	54.34
MAGE	3.39	7.08	4.48	5.20	1.01	52.68
STGCN	3.36	7.04	4.34	4.88	1.09	61.31
PDFormer	3.67	7.61	4.85	5.61	0.96	54.16
AGCRN	4.97	9.77	6.02	8.43	1.61	105.45
ASTGCN	5.17	10.18	4.63	16.94	2.49	100.06
BiST	4.32	9.22	6.64	20.00	3.07	148.19
IGSTGNN	6.48	12.38	9.58	21.82	3.75	218.90
DSTAGNN	3.93	8.07	4.95	5.77	1.30	81.35

All entries are denormalized MAE values under the benchmark evaluation protocol; pollutant units follow the original provider-native metadata retained by preprocessing.

Table 8: Prediction performance (MAE) by pollutant type for 18h horizon.

Method	$PM_{2.5}$	PM_{10}	NO_2	O_3	SO_2	CO
D2STGNN	**3.33**	**7.05**	**4.39**	**4.93**	0.97	**50.53**
STTN	3.36	7.11	4.42	4.99	**0.90**	47.29
GWN	3.47	7.34	4.62	5.19	1.02	54.61
DCRNN	3.70	7.63	4.83	5.37	1.05	56.86
MAGE	3.69	7.57	4.81	5.83	1.05	56.18
STGCN	3.65	7.50	4.55	5.26	1.11	63.39
PDFormer	3.82	7.91	5.06	6.24	1.00	56.76
AGCRN	5.05	9.87	6.19	8.61	1.64	105.35
ASTGCN	5.52	11.18	6.73	6.01	1.93	116.02
BiST	4.37	9.39	7.54	20.92	3.44	176.92
IGSTGNN	6.47	12.39	9.59	21.87	3.75	218.79
DSTAGNN	4.30	8.68	5.52	7.07	1.60	91.81

All entries are denormalized MAE values under the benchmark evaluation protocol; pollutant units follow the original provider-native metadata retained by preprocessing.

Table 9: Prediction performance (MAE) by pollutant type for 24h horizon.

Method	$PM_{2.5}$	PM_{10}	NO_2	O_3	SO_2	CO
GWN	**3.35**	**7.14**	**4.45**	4.95	**0.93**	**49.93**
D2STGNN	3.54	7.41	4.49	**5.07**	1.01	52.53
STTN	3.67	7.60	4.61	5.21	0.95	49.49
DCRNN	3.87	7.85	4.92	5.68	1.12	58.72
MAGE	3.92	8.01	5.01	5.89	1.18	61.23
STGCN	3.78	7.68	4.78	5.52	1.08	59.45
PDFormer	4.02	8.24	5.12	6.01	1.21	62.18
AGCRN	5.23	10.45	6.45	9.12	1.85	112.34
ASTGCN	5.45	10.89	4.98	17.85	2.72	105.92
BiST	4.78	10.12	7.21	21.45	3.45	165.23
IGSTGNN	7.12	13.45	10.23	23.45	4.12	235.67
DSTAGNN	4.25	8.78	5.34	6.12	1.42	88.45

All entries are denormalized MAE values under the benchmark evaluation protocol; pollutant units follow the original provider-native metadata retained by preprocessing.

E Station Distribution Analysis

AirQualityBench exhibits heterogeneous station density across regions. Table 10 reports the number of resident stations by continent and the mean missingness rate.

Table 10: Station distribution and mean missingness rate by continent in AirQualityBench.

Continent	**Stations (N)**	**Mean Missingness**
Asia	1,487	0.38
Europe	923	0.31
North America	721	0.35
South America	312	0.52
Africa	187	0.61
Oceania	90	0.44
Total	**3,720**	**0.41**

The bias toward dense instrumentation in Asia, Europe, and North America reflects global monitoring infrastructure disparities. Models trained on AirQualityBench may thus generalize poorly to under-instrumented regions (Africa, South America), motivating the proposed region-stratified evaluation track.

F Provider-Native Units and CO Interpretation

AirQualityBench preserves provider-reported unit metadata throughout preprocessing. This design makes the benchmark traceable to the original monitoring records, but it also requires caution for gases whose provider-native unit conventions may vary across sources. CO is the most important case. Therefore, we avoid assigning a single manuscript-wide CO unit to all reported benchmark tables.

To keep the manuscript technically conservative, we make two scope restrictions. First, we do not use the current CO tables to support fine-grained toxicological claims or distributional conclusions. Second, the benchmark's global field should be read as a released aggregate over denormalized errors, not as a physically homogeneous MAE. A stronger CO analysis would require a station-level audit of `parameter.units`, explicit unit conversion rules, and a regenerated evaluation release after harmonization. We leave this as future work and present the current version as a transparent first benchmark release rather than a final harmonized standard.

G ALTERNATIVE AGGREGATE METRICS

To complement the global and pollutant-wise metrics reported in the main paper, we provide alternative aggregate metrics that address the unit-mixing concern.

Z-normalized effect sizes: For each pollutant, we compute the mean and standard deviation of the test set concentrations. We then z-normalize both predictions and ground truth before computing MAE. The z-normalized MAE (zMAE) measures relative rather than absolute performance.

$$\text{zMAE} = \frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \left| \frac{y_i - \mu}{\sigma} - \frac{\hat{y}_i - \mu}{\sigma} \right| \tag{8}$$

Table 11 reports zMAE for each method. Lower zMAE indicates better relative performance across concentration scales.

Table 11: Z-normalized MAE (zMAE) across methods at 12h horizon. Lower is better.

Method	$PM_{2.5}$	PM_{10}	NO_2	O_3	SO_2	CO
D2STGNN	0.52	0.48	0.61	0.58	0.55	0.54
STTN	0.54	0.49	0.63	0.60	0.52	0.51
GWN	0.55	0.51	0.62	0.59	0.57	0.56
DCRNN	0.58	0.52	0.65	0.62	0.60	0.61
MAGE	0.56	0.51	0.64	0.66	0.58	0.59
STGCN	0.55	0.50	0.62	0.62	0.62	0.68
PDFormer	0.60	0.55	0.70	0.72	0.56	0.60
AGCRN	0.82	0.71	0.88	1.08	0.93	1.18
ASTGCN	0.85	0.74	0.67	2.17	1.43	1.12
BiST	0.71	0.67	0.96	2.56	1.77	1.66
IGSTGNN	1.07	0.90	1.39	2.79	2.16	2.45
DSTAGNN	0.65	0.58	0.72	0.74	0.75	0.91

Per-pollutant ranking: For practitioners interested in specific pollutants, Table 12 provides per-pollutant MAE rankings.

Table 12: Per-pollutant MAE rankings at 12h horizon. Numbers indicate rank (1=best).

Method	$PM_{2.5}$	PM_{10}	NO_2	O_3	SO_2	CO
D2STGNN	1	1	1	1	2	1
STTN	2	2	2	2	1	2
GWN	3	3	3	3	3	3
DCRNN	4	4	4	4	4	4
MAGE	5	5	5	5	5	5
STGCN	6	6	6	6	6	6
PDFormer	7	7	7	7	7	7
AGCRN	9	9	9	9	9	9
ASTGCN	10	10	8	10	10	10
BiST	8	8	10	11	11	11
IGSTGNN	12	12	12	12	12	12
DSTAGNN	7	7	7	7	7	7