BENCHMARKING PATCHTST AGAINST STATISTICAL AND MACHINE-LEARNING BASELINES FOR DAILY $PM_{2.5}$ FORECASTING IN TERMIZ, UZBEKISTAN

Din, Irfanud Ud[1, a)]
Raxmanova Oydina[2, b)]
Raxmanov Mirkomil[3*, c)]
Author Affiliations
[1]*Department of Computer Science, New Uzbekistan University, Tashkent, Uzbekistan*
[2]*Department of Computer Science, New Uzbekistan University, Tashkent, Uzbekistan*
[3]*Department of Artificial Intelligence, Tashkent University of Information Technologies, Tashkent,*
Uzbekistan
Author Emails
Corresponding author: irfan@newuu.uz
oydinajoraboyeva2604@gmail.com
mirkomilraxmanov@gmail.com

Abstract. *This study reassesses univariate Patch Time Series Transformer (PatchTST) forecasts of daily city-level $PM_{2.5}$ estimates for Termiz, Uzbekistan, using a controlled benchmark against persistence, annual seasonal naive, Ridge, XGBoost, and LSTM models. Archived chronological test evaluations cover 7-, 14-, and 30-day direct forecast horizons, with 14 nonoverlapping forecast origins per horizon and 98, 196, and 420 unique target dates, respectively. PatchTST achieved MAE values of 23.649, 20.546, and 20.140 μg/m³, and RMSE values of 32.595, 29.270, and 28.306 μg/m³ across these horizons. The corresponding R^2 values were 0.008, 0.205, and 0.121. The evaluation audit reports no duplicate target dates within model–horizon combinations. These results improve upon the earlier single-model diagnostic, but the study remains limited to univariate inputs and a small number of test origins. The contribution is a transparent, horizon-controlled local benchmark rather than a validated operational air-quality forecasting system.*

Keywords: *$PM_{2.5}$; PatchTST; Termiz; time-series forecasting; chronological evaluation; baseline comparison; forecast horizon.*

Annotatsiya. *Ushbu tadqiqotda O‘zbekistonning Termiz shahri uchun kunlik $PM_{2.5}$ konsentratsiyasining baholangan qiymatlarini bir o‘zgaruvchili Patch Time Series Transformer (PatchTST) modeli yordamida prognozlash natijalari qayta baholandi. Bunda PatchTST modeli Persistence (oxirgi kuzatilgan qiymatni saqlash), yillik mavsumiy sodda prognozlash (Annual Seasonal Naive), Ridge regressiyasi, XGBoost va LSTM modellari bilan nazorat qilinadigan qiyosiy tahlil asosida solishtirildi. Arxivlangan xronologik test natijalari 7, 14 va 30 kunlik bevosita prognozlash gorizontlarini qamrab oldi. Har bir prognozlash gorizonti uchun o‘zaro kesishmaydigan 14 ta prognozlash boshlang‘ich nuqtasi tanlandi va mos ravishda 98, 196 hamda 420 ta takrorlanmaydigan maqsadli sana baholandi.*

Kalit so‘zlar: *$PM_{2.5}$; PatchTST; Termiz; vaqt qatorlarini prognozlash; xronologik baholash; bazaviy modellar bilan taqqoslash; prognozlash gorizonti.*

Аннотация. *В данном исследовании проведена повторная оценка результатов прогнозирования ежедневных оценочных концентраций $PM_{2.5}$ на уровне города Термез (Узбекистан) с использованием одномерной модели Patch Time Series Transformer (PatchTST). Оценка выполнена в рамках контролируемого сравнительного анализа с моделями Persistence (сохранение последнего наблюдаемого значения), Annual Seasonal Naive (сезонный наивный прогноз с годовой периодичностью), Ridge-регрессией, XGBoost и LSTM. Архивные результаты хронологического тестирования охватывают горизонты прямого прогнозирования на 7, 14 и 30 дней. Для каждого горизонта использовалось 14 непересекающихся начальных точек прогнозирования, что соответствует 98, 196 и 420 уникальным целевым датам соответственно.*

Ключевые слова: *$PM_{2.5}$; PatchTST; Термез; прогнозирование временных рядов; хронологическая оценка; сравнение с базовыми моделями; горизонт прогнозирования.*

INTRODUCTION

Fine particulate matter ($PM_{2.5}$) is an important indicator of ambient air quality. The WHO 2021 guidelines provide health-based reference concentrations of 5 μg/m³ annually and 15 μg/m³ over 24 hours [1]. Forecasting is challenging because the observed signal combines persistence, seasonality, abrupt events, and exogenous influences. Transformer architectures [2] and PatchTST [3] offer a flexible representation of long time series, but their advantages must be tested against credible baselines rather than inferred from model complexity.

This paper updates a previously reported single-model evaluation using new archived predictions, fixed forecast origins, explicit target-date auditing, and comparisons at 7-, 14-, and 30-day horizons. Its contribution is a local benchmark and

evaluation audit, not a claim of a new Transformer architecture or a validated operational air-quality warning system.

MATERIALS AND METHODS

Study data and quality control

The analytical source is identified in the experiment metadata as "termiz_pm25_10_rows.csv" and described in the prior manuscript as originating from the World Bank Daily City-Level Air Pollution $PM_{2.5}$ Estimates dataset [4]. These are city-level estimates, not necessarily direct readings from a ground station. The archive does not contain the original source CSV, retrieval timestamp, license statement, identifier audit, or the original raw values; therefore independent provenance and missing-value verification remain outstanding.

TABLE 1. Variables in the analytical dataset.

Variable	Role / Description
wb_region	World Bank region
nam_0	**Country name**
adm0_wb_a3	Country administrative code
pm25_iso3	ISO3 code associated with $PM_{2.5}$ record
adm1cd_c	First-level administrative code
Geonameid	Geographic identifier
City	**City name; filtered to Termiz**
Latitude	City latitude
Longitude	City longitude
Population	Population field supplied with the source data
Date	Daily observation date
pm25	**Target $PM_{2.5}$ concentration (µg/m³)**

The archived quality report lists 2,797 raw rows, 2,891 calendar dates, and 94 missing $PM_{2.5}$ values after daily reindexing (3.25% of calendar dates). The report states a forward-fill limit of two days and no future-value interpolation. It does not establish how any remaining gaps were treated or whether all training transformations were fitted strictly within training; these are limits of the present reproducibility audit.

TABLE 2. Chronological split and supervised windows

Partition	Start	End	Calendar days
Training	2018-10-01	2024-04-14	2023
Validation	2024-04-15	2025-06-22	434
Test	2025-06-23	2026-08-30	434

A 90-day lookback forms the input to each direct forecasting task. Patch length is 15 days and stride is 5, implying 16 patches without end padding: $\text{floor}((90-15)/5)+1 = 16$. The archived configuration does not independently estab-

lish the effective padding/window count at every partition boundary. Historical context preceding a split may be used as model input, but no target labels from later splits may enter model fitting or preprocessing.

Standardization was fitted exclusively on the training subset. For observation x_i, the standardized value z_i was computed from the training-set mean μ_{train} and standard deviation σ_{train}:

$$z_i = (x_i - \mu_{train}) / \sigma_{train} \quad (1)$$

Predictions were transformed back to μg/m³ before reporting performance. Fitting the scaler only on the training period prevents information from validation and test observations from entering preprocessing[8].

PatchTST Architecture

The forecasting task was formulated as direct multi-step prediction: the previous 90 daily $PM_{2.5}$ values formed the context window and the next 30 daily values formed the prediction target. Thus, the model estimates a future sequence from a fixed historical window:

$$[y_{t-89}, \ldots, y_t] \rightarrow [\hat{y}_{t+1}, \ldots, \hat{y}_{t+30}] \quad (2)$$

Each 90-day input was divided into overlapping patches of length 15 with stride 5. These patch tokens were embedded and passed through a Transformer encoder[9]. In the present univariate implementation, $PM_{2.5}$ is the single input channel. The model therefore learns temporal dependence in $PM_{2.5}$ itself; it does not establish causal effects of meteorology, traffic, industry, or other external drivers.

TABLE 3. PatchTST hyperparameters used in the experiment.

Parameter	Value
Input channels	1
Context length	90 days
Prediction length	30 days
Patch length	15
Patch stride	5
Model dimension (d_model)	128
Transformer encoder layers	3
Attention heads	4
Feed-forward dimension	256
Dropout / head dropout	0.10 / 0.10
Batch size	32
Maximum epochs	50
Learning rate	1×10^{-4}
Optimizer	AdamW
Training loss	Mean squared error

Model configuration and comparator methods

The saved Hugging Face Transformers PatchTSTForPrediction configurations (Transformers version 4.57.6) specify one input channel, context length 90, d_model 64, two encoder layers, four attention heads, feed-forward dimension 128, patch length 15, patch stride 5, dropout 0.15, attention dropout 0.10, and MSE loss. Separate checkpoints exist for horizons 7, 14, and 30. This replaces the earlier manuscript's 128-dimensional, three-layer description, which applies to the earlier experiment rather than the archived benchmark. The compared methods are persistence, 365-day seasonal naive, Ridge, XGBoost, and LSTM. Model-specific fitting parameters for these comparators are not fully documented in the supplied archives[9].

Evaluation and statistical analysis

The archive identifies a random seed of 2026 and an evaluation stride of 30. For each horizon, 14 forecast origins are used, with nonoverlapping target blocks and one prediction per target date within each model/horizon. Thus n=98, 196, and 420 for horizons 7, 14, and 30, respectively. These counts refer to evaluated forecast targets, not to all calendar days in the 434-day test partition. The horizons cover different date sets, so cross-horizon MAE differences are descriptive and cannot be attributed solely to horizon length[10].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \acute{y}_i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \acute{y}_i)^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \acute{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \acute{y}_i}{y_i}\right| \quad (6)$$

The archived relative-error measure is $MAPE$, subject to a defined zero-denominator convention[11]. This is not interchangeable with the MAPE reported in the earlier manuscript. Bias is reported from the archived evaluation and its sign convention should be confirmed in the producing code. Exploratory paired bootstrap intervals resample forecast origins; dependence across nearby origins may make them optimistic.

RESULTS

TABLE 4. Benchmark performance on audited test targets

Model	H (days)	N	MAE	RMSE	R^2	sMAPE (%)
PatchTST	**7**	**98**	**23.649**	**32.595**	**0.008**	**31.75**
Ridge	7	98	24.127	34.674	-0.123	32.74
XGBoost	7	98	27.875	39.214	-0.436	38.30
LSTM	7	98	28.934	41.075	-0.575	39.97

Persistence	7	98	31.568	46.885	-1.052	37.63
Seasonal-Naive365	7	98	31.642	43.636	-0.778	46.02
PatchTST	**14**	**196**	**20.546**	**29.270**	**0.205**	**28.15**
Ridge	14	196	24.254	35.407	-0.164	32.68
XGBoost	14	196	27.070	38.971	-0.410	37.10
LSTM	14	196	28.580	41.224	-0.577	39.55
Persistence	14	196	30.399	44.113	-0.806	35.86
Seasonal-Naive365	14	196	31.198	43.878	-0.787	44.77
PatchTST	**30**	**420**	**20.140**	**28.306**	**0.121**	**27.66**
Ridge	30	420	23.354	32.703	-0.173	32.67
XGBoost	30	420	28.235	38.344	-0.612	40.17
LSTM	30	420	29.167	39.909	-0.746	41.67
Seasonal-Naive365	30	420	29.173	39.583	-0.718	42.84
Persistence	30	420	30.749	45.227	-1.243	35.84

All MAE and RMSE values are in μg/m³. PatchTST has the lowest MAE and RMSE in each horizon-specific comparison in the archived results. For the 30-day task, PatchTST MAE is 20.140 μg/m³ compared with 23.354 for Ridge, 28.235 for XGBoost, 29.167 for LSTM, 29.173 for seasonal naive, and 30.749 for persistence. The R^2 values for PatchTST are positive but modest, indicating limited explanatory skill and leaving substantial unexplained variability[12].

Date-alignment audit

The archived evaluation audit reports zero duplicated target dates within every model/horizon, 14 unique origins for each model/horizon, and 98/196/420 unique target dates. Test targets run from 2025-06-23 to 2026-07-24 (7-day), 2026-07-31 (14-day), and 2026-08-16 (30-day). The archived analysis additionally finds 98 common (date, lead) pairs across all three horizons. Accordingly, pooled overlapping-window metrics from the earlier manuscript must not be combined with these unique-date results[13].

TABLE 5. Paired-origin uncertainty

Horizon	Comparator	MAE difference (baseline − PatchTST)	Exploratory 95% interval
7	Persistence	7.919	[-5.564, 25.148]
7	Ridge	0.478	[-3.953, 5.029]
7	SeasonalNaive365	7.993	[1.696, 14.551]
14	Persistence	9.854	[-2.378, 25.985]
14	Ridge	3.708	[-0.166, 8.251]
14	SeasonalNaive365	10.652	[5.663, 17.145]
30	Persistence	10.609	[1.304, 26.200]

30	Ridge	3.214	[1.062, 5.666]
30	SeasonalNaive365	9.033	[6.156, 12.411]

Positive MAE differences favor PatchTST. Some 7- and 14-day intervals cross zero, especially for Ridge and persistence. These intervals are exploratory, not definitive significance tests: only 14 origins are available, and origin-resampling does not fully account for serial dependence[14].

Visual diagnostics

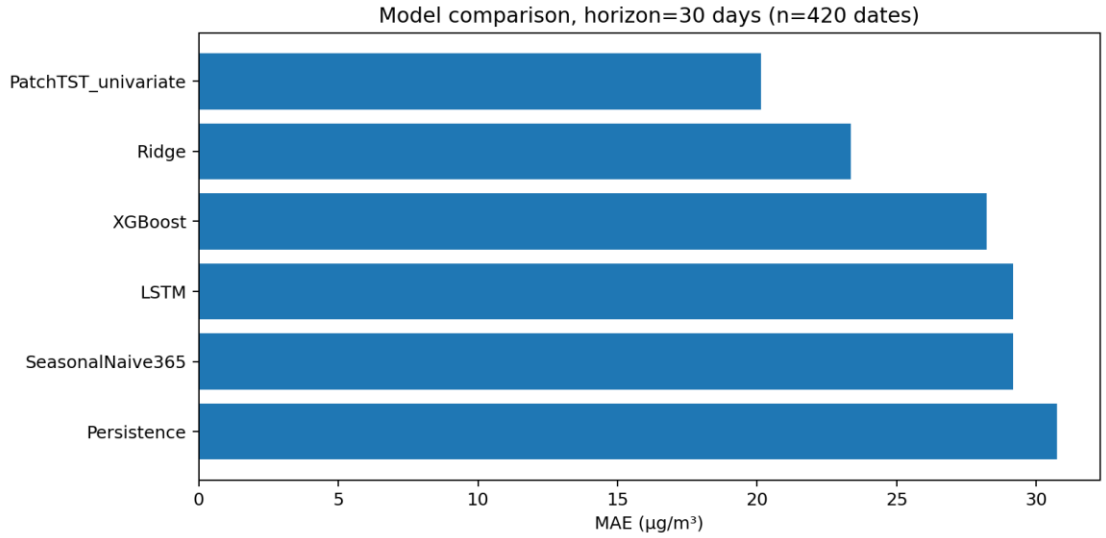

FIGURE 1. MAE comparison across models for the 30-day horizon.

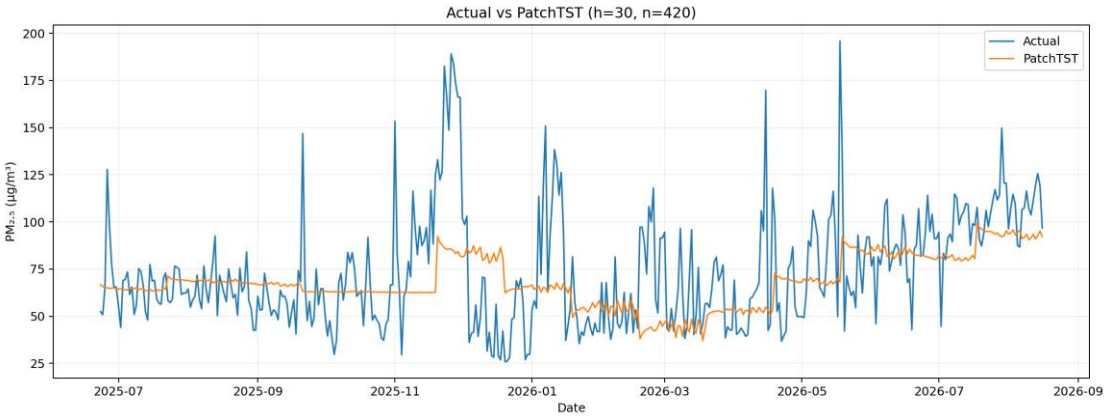

FIGURE 2. Archived actual-versus-PatchTST series for the 30-day task.

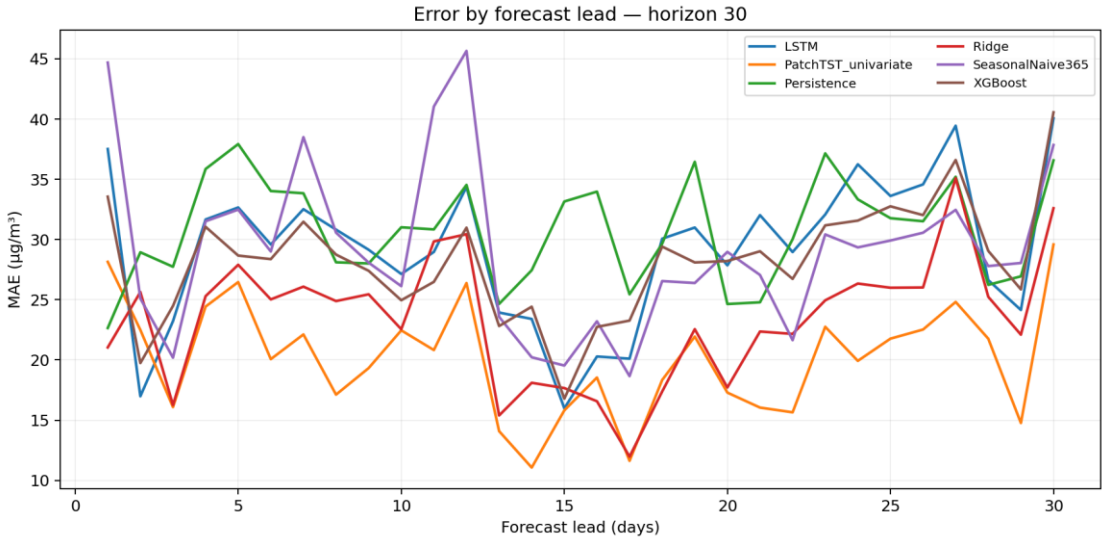

FIGURE 3. Forecast-lead MAE for the 30-day task.

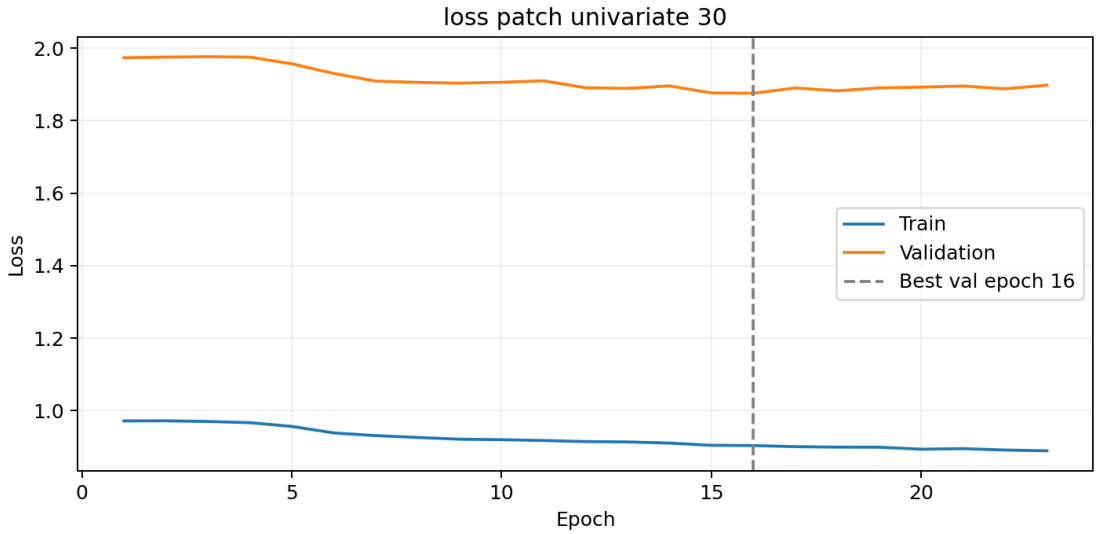

FIGURE 4. Archived training and validation loss for the 30-day PatchTST model.

These figures are reproduced from the supplied analysis archive; their underlying prediction files are available in the archive. The actual epoch-selection rule, early-stopping patience, optimizer weight decay, hardware, and full training environment are not independently established by the saved charts and configurations alone.

DISCUSSION

The updated benchmark materially changes the interpretation of the earlier negative-R^2 PatchTST report. Under the newly archived date-aligned protocol, PatchTST achieves positive R^2 and lower MAE than five baseline models across all three reported tasks. These results support relative improvement within this specific benchmark, not a general claim of deployment readiness. The test date sets vary by horizon; even a seemingly better 30-day MAE than 7-day MAE is not evidence that longer forecasts are intrinsically easier. Missing meteorological, dust, and emission covariates may limit performance, but this univariate study cannot establish their causal roles[15].

The former recursive 365-day trajectory, which converged to a near-constant 100.4 μg/m³, is not included as a validated future forecast in the revised results. It is a diagnostic of recursive collapse in the earlier configuration. Its minimum and maximum point forecasts are not uncertainty bounds. The present archive contains direct 7-, 14-, and 30-day evaluations, not a new independently validated 365-day forecast[16].

Limitations and reproducibility

The supplied archive contains model checkpoints, model configuration files, prediction records, evaluation audits, metrics, loss logs, and figures. However, the original dataset, exact download/version and license, full training scripts, package environment beyond the recorded Transformers version, comparator hyperparameters, and complete gap-handling implementation were not supplied. Therefore, numerical benchmark artifacts can be re-analysed but end-to-end training and data provenance cannot yet be independently reproduced. Future work should publish these materials, implement rolling-origin validation with more independent origins, and test weather-aware multivariate configurations[17][18].

CONCLUSION

A date-aligned benchmark of daily Termiz $PM_{2.5}$ estimates compared PatchTST with five forecasting baselines at 7-, 14-, and 30-day horizons. In the archived evaluation, PatchTST achieved MAE values of 23.649, 20.546, and 20.140 μg/m³, respectively, with positive but limited R^2. Unique-date auditing eliminated ambiguity from overlapping evaluation windows and established a consistent 14-origin protocol. The model ranked first among the six compared methods by MAE on the archived tests, while bootstrap uncertainty, differing horizon-specific date coverage, and incomplete source/training provenance restrict stronger conclusions. The findings support PatchTST as a promising benchmark for this dataset, but not yet as a validated operational forecast system.

DATA AND CODE AVAILABILITY

The dataset used in this study contains daily $PM_{2.5}$ concentration measurements for Termiz, Uzbekistan. The data supporting the findings of this research are publicly available at [https://datacatalog.worldbank.org/search/dataset/0067162/daily-city-level-air-pollution-pm2-5-estimates-2018-2026].

The source code for data preprocessing, model training, hyperparameter optimization, and performance evaluation, including the PatchTST model and benchmark forecasting methods, is publicly available at [https://github.com/Mirkomil999/Air_pulution_Termiz.git].

The repository also contains the configuration files and instructions required to reproduce the experimental results presented in this study.

REFERENCES

1. World Health Organization, WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide (World Health Organization, Geneva, 2021).

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems 30 (2017).

2. Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," in International Conference on Learning Representations (ICLR) (2023).

3. World Bank, "Daily City-Level Air Pollution $PM_{2.5}$ Estimates, 2018–2026," World Bank Data Catalog, dataset 0067162 (metadata accessed 2026).

4. S. Hochreiter and J. Schmidhuber, Neural Computation 9, 1735–1780 (1997).

5. B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, International Journal of Forecasting 37, 1748–1764 (2021).

6. J. Jitsakul, "Fine Particulate Matter ($PM_{2.5}$) Concentration Prediction in an Urban Area of Thailand Using Deep Learning Models," *International Journal of Technology*, 2026. This is particularly valuable for your paper because the study explicitly compares PatchTST, Transformer, CNN-LSTM, recurrent models, statistical models, and persistence baselines using 24,840 hourly observations.

7. W. Kristjanpoller, A. Lopez, F. Mora, and M. C. Minutolo, "$PM_{2.5}$ forecasting: A comparative study of transfer learning with transformer models and panel data approaches for country-level predictions," *Expert Systems with Applications*, vol. 317, Art. 131971, 2026. DOI: 10.1016/j.eswa.2026.131971. This supports discussion of Transformer transfer learning and generalization across locations.

8. "Evaluating deep learning time series models for $PM_{2.5}$ forecasting across diverse horizons," *iScience*, vol. 29, no. 2, Art. 114770, 2026. DOI: 10.1016/j.isci.2026.114770. The paper compares LSTM, CNN-LSTM, Transformer, and Transformer-LSTM and reports that meteorological variables such as temperature, pressure, precipitation, and wind speed materially improve forecasting. This is especially relevant to explaining why your current univariate PatchTST model has limited generalization.

9. "An Adaptive Spatiotemporal Graph Transformer for Multi-Site $PM_{2.5}$ Multi-Step Forecasting with Non-stationary and Sparse-Aware Method," *Journal of Environmental Management*, vol. 414, Art. 130515, 2026. DOI: 10.1016/j.jenvman.2026.130515. This provides recent evidence for addressing non-stationarity, spatial dependencies, and multi-step forecasting using Transformer architectures.

10. "Adaptive Graph Neural Network–Transformer Model for High-Resolution $PM_{2.5}$ Forecasting and Spatial Extrapolation," *Environmental Modelling & Software*, vol. 196, Art. 106771, 2026. DOI: 10.1016/j.envsoft.2025.106771. The study integrates graph neural networks, temporal convolutions, and Transformers and uses meteorological and geographic information.

11. Y. Cheng and C. Zhang, "Forecasting of $PM_{2.5}$ concentration based on variational mode decomposition and deep learning," *Scientific Reports*, vol. 16, Art. 25647, 2026. The proposed method combines VMD, TCN, BiLSTM, and attention to address nonlinear and non-stationary $PM_{2.5}$ behavior.

12. D. Li, J. Jin, G. Wang, M. Pang, W. Zhang, and H. Liao, "A Transformer-based agent model of GEOS-Chem v14.2.2 for informative prediction of $PM_{2.5}$ and O_3 levels to future emission scenarios: TGEOS v1.0," *Geoscientific Model Development*, vol. 19, pp. 1703–, 2026. DOI: 10.5194/gmd-19-1703-2026. This is useful for discussing Transformer models beyond purely historical time-series forecasting and toward emission-scenario modeling.

13. Y. Yu et al., "Boosting $PM_{2.5}$ short-term prediction by coupling graph attention and Koopman networks," *Computational Urban Science*, vol. 6, Art. 16, 2026. This provides a recent alternative deep-learning direction for handling the complex nonlinear behavior of $PM_{2.5}$.

14. "Data Augmentation Strategies for Improved $PM_{2.5}$ Forecasting Using Transformer Architectures," *Atmosphere*, vol. 16, no. 2, Art. 127, 2025. DOI: 10.3390/atmos16020127. This one is highly relevant to your results because it investigates data imbalance and extreme $PM_{2.5}$ events, which your current model also struggles to reproduce.

15. R. Zhang and N. Awang, "A novel ST-iTransformer model for spatio-temporal ambient air pollution forecasting," *Journal of Big Data*, vol. 12, Art. 101, 2025. DOI: 10.1186/s40537-025-01150-5. This is useful for the Related Work section on newer spatiotemporal Transformer architectures.

16. J. Bernacki, "A hybrid transformer and symbolic regression model for weather-dependent particulate matter forecasting in air quality management," *Theoretical and Applied Climatology*, vol. 156, Art. 662, 2025. Published 19 November 2025. This supports your proposed move from univariate $PM_{2.5}$ forecasting toward weather-dependent modeling.

17. "Development of a decomposition-optimization-transformer hybrid model for spatiotemporal forecasting of $PM_{2.5}$ air pollution in Chinese cities: A case study," *Ecotoxicology and Environmental Safety*, vol. 305, Art. 119209, 2025. DOI: 10.1016/j.ecoenv.2025.119209. The model combines decomposition, optimization, and Transformer components and reports strong $PM_{2.5}$ forecasting performance across multiple monitoring stations.