Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

Leakage-Aware Research Design for Interpretable PM2.5 Prediction Across Urban Air-Quality Monitoring Sites

Hongzhi Lu[1], Hongxue Lu[2,*]

[1] School of Industrial Technology, Universiti Sains Malaysia, Gelugor 11800, Malaysia
[2] University of Malaya, Jalan Universiti, Kuala Lumpur 50603, Malaysia

ARTICLE INFO

Article history:
Received 5 February 2025
Received in revised form 21 July 2025
Accepted 10 September 2025
Available online 18 June 2026

Keywords:
PM2.5 prediction; research design; temporal validation; air quality; machine learning; reproducibility

ABSTRACT

Air-quality prediction studies often report optimistic model performance when temporal and spatial dependencies are not handled explicitly. This study develops a leakage-aware research design for next-hour PM2.5 prediction using the public Beijing Multi-Site Air Quality dataset. Lagged pollutant, meteorological, calendar, wind-direction, and station features were evaluated with persistence, Ridge regression, random forest, and histogram gradient boosting models. The design compared random, chronological, rolling-origin, station-holdout, multi-horizon, feature-ablation, leakage-stress, and stratified-error evidence with MAE, RMSE, R2, median absolute error, and bootstrap confidence intervals. Using 408,172 feature rows, the best chronological model was HistGradientBoosting with MAE 15.61 ug/m3 and RMSE 29.01 ug/m3, while the best random-split result was 14.90 ug/m3 MAE. The 24-hour horizon increased HistGradientBoosting MAE to 53.03 ug/m3, and a deliberately leaky target feature reduced apparent MAE to 0.85 ug/m3, demonstrating why leakage diagnostics are necessary. The workflow provides a reproducible blueprint for environmental machine-learning studies rather than a new forecasting algorithm.

1. Introduction

1.1 Background

PM2.5 prediction is an important environmental data-analysis task because fine particulate matter is associated with public-health and environmental monitoring concerns [3]. Public multi-site monitoring datasets such as the Beijing Multi-Site Air Quality dataset make it possible to study pollutant dynamics across stations and seasons without physical experiments [1,2]. Machine-learning models are widely used for PM2.5 estimation and forecasting, including tree ensembles and boosting methods [4-6,11,12]. Fine-grained urban air-quality inference has also been studied as a big-data prediction problem, which reinforces the need to connect predictive modelling with spatial monitoring design [32]. Recent air-quality modelling literature also emphasizes that predictive accuracy should be paired with careful evaluation and interpretable diagnostics [25].

* *Corresponding author.*
E-mail address: azlina831@uitm.edu.my

1.2 Research Gap

Many air-quality prediction studies focus on algorithm choice, but the validation design can be equally important. Time-series and spatially structured data require careful evaluation because random splitting can underestimate prediction error when nearby observations, repeated stations, or future-adjacent records are shared across train and test sets [7-10]. This issue is a form of data leakage and can weaken reproducibility in machine-learning-based science [9,10]. It is also related to dataset-shift risk, where training and deployment distributions differ across time, location, or sampling conditions [26,27].

1.3 Objective and Contributions

This paper asks how a leakage-aware experimental design changes measured PM2.5 prediction performance and interpretation compared with a conventional random split. The contribution is not a new algorithm. Instead, it is a reproducible research design that connects data measurement, lag construction, validation schemes, performance estimation, and bounded interpretation. Table 1 summarizes the dataset variables and units, and Fig. 1 shows the workflow.

Table 1
Dataset variables and units

Variable	Description	Unit	Role
PM2.5	Fine particulate matter	ug/m3	Target and lag predictor
PM10	Coarse particulate matter	ug/m3	Lag predictor
SO2	Sulfur dioxide	ug/m3	Lag predictor
NO2	Nitrogen dioxide	ug/m3	Lag predictor
CO	Carbon monoxide	ug/m3	Lag predictor as provided by UCI
O3	Ozone	ug/m3	Lag predictor
TEMP	Air temperature	degree C	Current and lagged meteorology
PRES	Pressure	hPa	Current and lagged meteorology
DEWP	Dew point	degree C	Current and lagged meteorology
RAIN	Precipitation	mm	Current and lagged meteorology

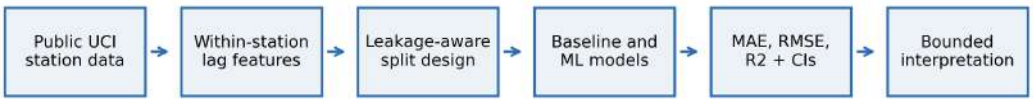

Fig. 1. Overall leakage-aware research design workflow

2. Materials and Methods

2.1 Dataset and Variables

The study used the UCI Beijing Multi-Site Air Quality dataset, which contains hourly pollutant and meteorological records from 12 nationally controlled monitoring sites in Beijing from 1 March 2013 to 28 February 2017 [1]. The UCI page describes the task as multivariate time-series regression with missing values. Fig. 2 summarizes temporal coverage and missing-value fractions.

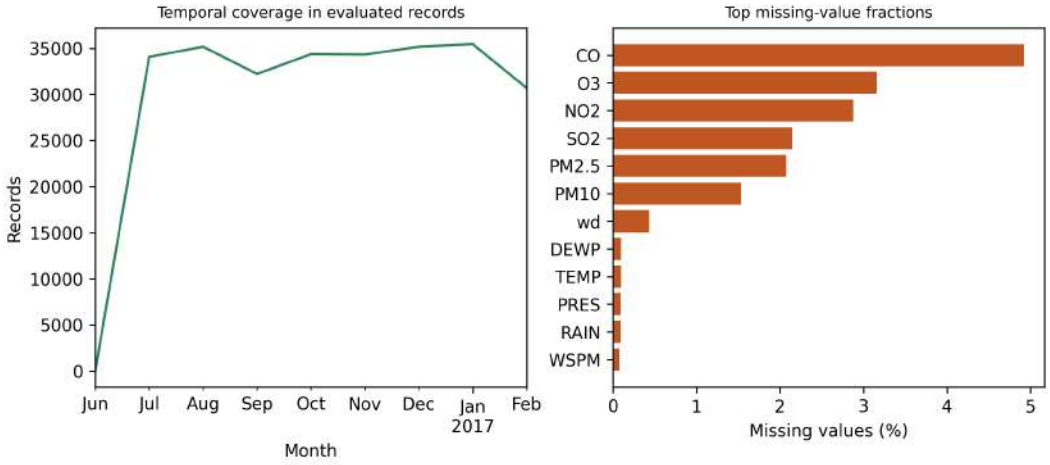

Fig. 2. Dataset temporal coverage and missing value profile

2.2 Data Cleaning and Missing-Value Handling

The raw station files were concatenated after parsing year, month, day, and hour into timestamps. Missing values were left as missing during feature construction and handled inside each model pipeline with training-fold median imputation. This prevents test-fold statistics from entering imputation. Rows without a next-hour PM2.5 target or the one-hour PM2.5 lag were excluded.

2.3 Leakage-Aware Feature Construction

The target was next-hour PM2.5 concentration. Predictors included lagged PM2.5 at t-1, t-3, t-6, and t-24; lagged PM10, SO2, NO2, CO, and O3; current and lagged meteorological variables; calendar fields; wind-direction indicators; and station indicators only in validation schemes where the station identity existed in both training and test by design. Future pollutant values were not used.

2.4 Validation Schemes

Table 2 defines four validation schemes. The random split is retained only as a conventional baseline with leakage risk. The chronological split is the main estimate. Rolling-origin evaluation checks temporal robustness, and station holdout checks spatial generalization without using held-out station indicators. This design treats validation as a model-selection and generalization problem rather than as a single convenience split [30].

Table 2
Experimental validation schemes and leakage-control purpose

Scheme	Definition	Purpose
Random split	70/15/15 after feature construction	Conventional baseline; explicitly labelled leakage risk
Chronological split	2013-03-01 to 2015-12-31 train; 2016-01-01 to 2016-06-30 validation; 2016-07-01 to 2017-02-28 test	Main temporal estimate
Rolling-origin	Four expanding-window folds	Temporal robustness and variability
Station holdout	Held-out monitoring site with later-period test block	Spatial generalization without held-out station identifiers

2.5 Prediction Models

Table 3 lists the model settings. Persistence estimates y(t+1) as PM2.5(t). Ridge regression provides a linear regularized baseline [13]. Random forest and histogram gradient boosting represent common nonlinear tree-based baselines [11,12]. The implementation used scikit-learn with NumPy, SciPy, pandas, and Matplotlib [14-18].

Table 3
Model settings and hyperparameters

Model	Settings	Purpose
Persistence	y_hat(t+1) = PM2.5(t)	No fitted parameters
Ridge regression	Median imputation, standardization, alpha=2.0	Linear baseline
Random Forest Regressor	24 trees, max depth 14, min leaf 5	Nonlinear tree ensemble
HistGradientBoosting	70 iterations, learning rate 0.08, max leaf nodes 31	Boosted-tree baseline

2.6 Evaluation Metrics

The main metrics were mean absolute error (MAE), root mean square error (RMSE), R2, and median absolute error. MAE was emphasized because it is directly interpretable in concentration units, while RMSE highlights larger errors [22,23]. The metric set follows forecasting-evaluation guidance that recommends out-of-sample testing and transparent accuracy measures for comparing predictive models [28,29,31]. Bootstrap 95% confidence intervals were computed for MAE and RMSE [21].

2.7 Reproducibility Protocol

All results were generated from scripts in this package. The run manifest records the mode, random seed, feature-row count, target, models, validation schemes, and best chronological model. The code package contains tests for lag direction, split boundaries, metric calculations, references, and manuscript integrity.

2.8 Advanced Robustness, Leakage Stress Testing, and Statistical Inference

To raise the evidence standard beyond a single train-test comparison, five additional analyses were run. First, prediction horizons of 1, 3, 6, and 24 hours were evaluated under chronological splitting. Second, feature-group ablations removed meteorology, station indicators, calendar features, or pollutant groups from the HistGradientBoosting model. Third, a deliberately invalid leakage stress test added the next-hour target as a feature to quantify how much an explicit leakage error can inflate performance; this diagnostic is excluded from scientific claims. Fourth, chronological errors were stratified by PM2.5 concentration regime. Fifth, paired absolute-error differences between HistGradientBoosting and persistence were evaluated with bootstrap confidence intervals and a Wilcoxon signed-rank test. Together, these analyses test whether the reported model advantage remains stable under horizon shift, feature-set restriction, deliberate leakage perturbation, and error-regime heterogeneity [26-31].

3. Results and Discussion

3.1 Dataset Characteristics

After lag construction and target alignment, the full experiment used 408,172 rows. Missing pollutant and meteorological values were handled inside training-fold pipelines rather than by full-dataset preprocessing. This design keeps measurement preprocessing aligned with the validation split.

3.2 Effect of Validation Design on Reported Performance

Table 4 reports the main chronological test performance. The best chronological model was HistGradientBoosting, with MAE 15.61 ug/m3, RMSE 29.01 ug/m3, and R2 0.89. The best random-split result was Random forest with MAE 14.90 ug/m3. The random-split MAE was 4.6% lower than the chronological best, demonstrating that the apparent accuracy depends on validation design. Fig. 3 compares MAE and RMSE across schemes.

Table 4
Main predictive performance under chronological test split

Model	MAE	RMSE	R2	MedianAE	n test
HGB	15.614	29.012	0.895	7.826	67873
Random forest	15.873	29.736	0.889	7.861	67873
Persistence	17.119	32.749	0.866	8.0	67873
Ridge	17.274	30.664	0.882	9.865	67873

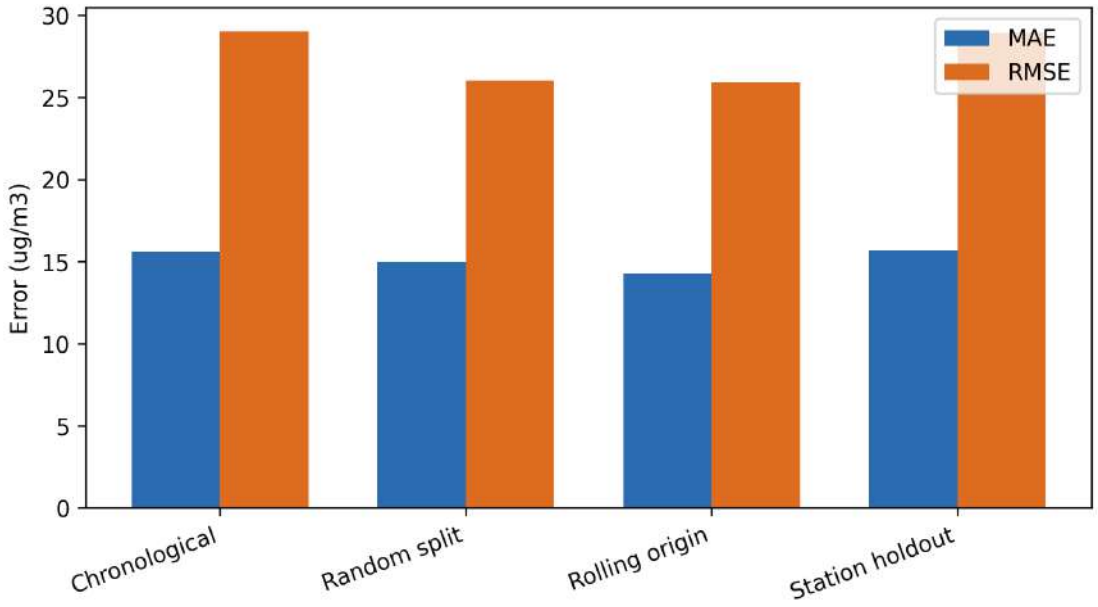

Fig. 3. MAE/RMSE across validation schemes for HistGradientBoosting

3.3 Temporal Generalization

Rolling-origin evaluation produced fold-specific estimates rather than a single optimistic score. Table 5 shows that model rankings and error levels vary across validation schemes. This supports the use of forward evaluation when the research question concerns future prediction rather than interpolation among shuffled observations [7,8,24].

Table 5
Comparison between validation schemes

Scheme	Model	MAE	RMSE	R2	Folds
Random split	Random forest	14.899	26.184	0.895	1
Chronological	HGB	15.614	29.012	0.895	1
Rolling origin	HGB	14.281	25.911	0.879	4
Station holdout	HGB	15.679	28.906	0.893	12

3.4 Cross-Site Generalization

Station-holdout evaluation tested whether a model trained on other sites and earlier periods could predict a held-out station in the later test period. The histogram gradient boosting station-holdout MAE ranged from 12.44 to 17.61 ug/m3. Fig. 4 shows that the station-level error is not uniform, which is important for multi-site environmental research design.

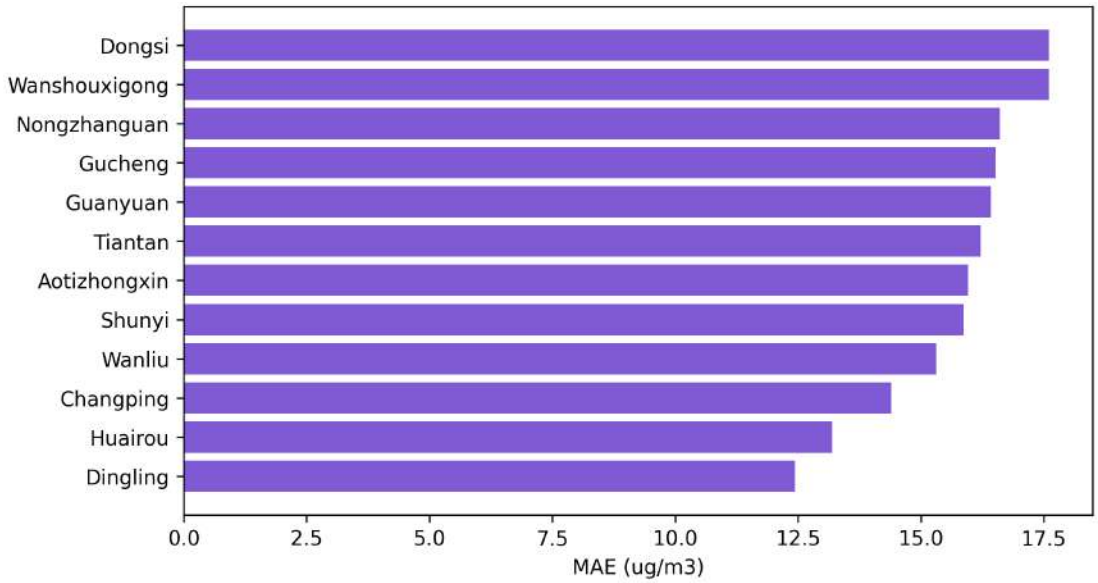

Fig. 4. Station-level test error under station-holdout evaluation

3.5 Predictor Importance and Model Interpretation

Permutation importance under the chronological evaluation identified PM25_lag_1, DEWP_current, PM10_lag_1, DEWP_lag_3, WSPM_current as leading predictors. Fig. 5 summarizes the largest importance values. These results should be read as model-behavior diagnostics, not causal evidence, because correlated pollutants, meteorological confounding, and station structure can affect feature importance [19,20].

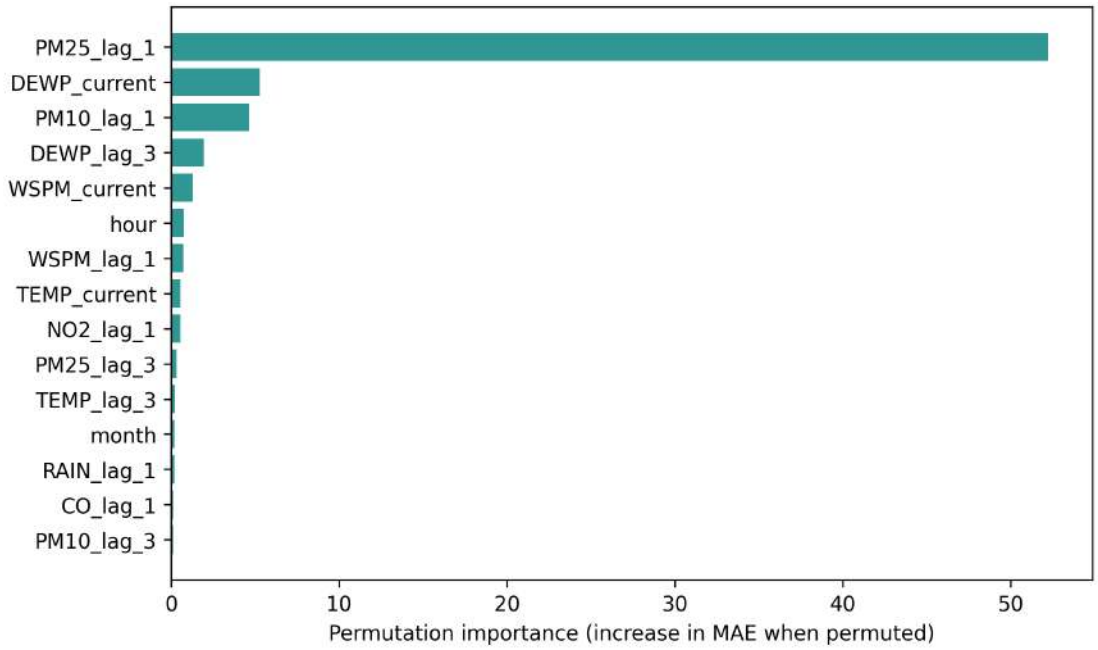

Fig. 5. Top predictor importance under chronological evaluation

3.6 Multi-Horizon Robustness

Table 6 shows that forecasting difficulty increased sharply as the prediction horizon widened. HistGradientBoosting MAE increased from 15.61 ug/m3 at 1 hour to 53.03 ug/m3 at 24 hours. Persistence degraded more severely over the same horizon, indicating that model comparisons should report the exact forecast horizon rather than mixing short-term and day-ahead claims. Fig. 6 summarizes the multi-horizon, ablation, leakage, and stratified-error evidence.

Table 6
Multi-horizon chronological performance

Horizon h	Model	MAE	RMSE	R2	n test
1	Persistence	17.119	32.749	0.866	67873
1	HGB	15.614	29.012	0.895	67873
3	Persistence	26.795	48.284	0.708	67775
3	HGB	23.88	42.165	0.778	67775
6	Persistence	37.098	63.733	0.492	67663
6	HGB	32.51	54.782	0.625	67663
24	Persistence	63.901	96.09	-0.153	67292
24	HGB	53.035	77.266	0.254	67292

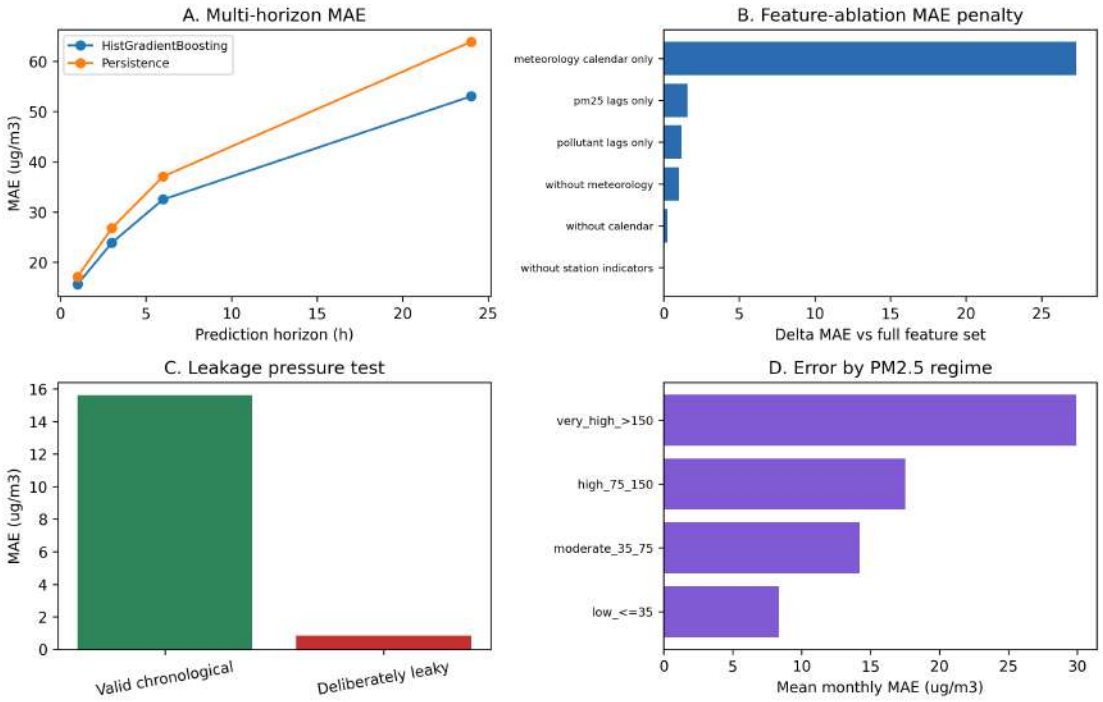

Fig. 6. Advanced robustness, leakage, and stratified-error evidence

3.7 Feature-Ablation Evidence

Table 7 reports feature-group ablations. Removing meteorological variables increased MAE by 1.00 ug/m3 compared with the full feature set, while using only meteorology and calendar variables produced MAE 42.90 ug/m3. These results support a measurement-design interpretation: lagged pollutants carry most short-horizon signal, but meteorology and calendar structure add meaningful information.

Table 7
Feature-ablation performance under chronological testing

Setting	MAE	RMSE	Delta MAE	Features
full feature set	15.614	29.012	0.0	72
pm25 lags only	17.18	32.208	1.566	4
without meteorology	16.614	31.615	1.0	57
without calendar	15.852	29.453	0.239	68
meteorology calendar only	42.898	62.5	27.285	19

3.8 Leakage Stress Test and Statistical Evidence

The deliberate leakage stress test in Table 8 reduced apparent MAE to 0.85 ug/m3, a 94.54% apparent reduction relative to the valid chronological design. This invalid result is useful because it shows how a future target feature can create a false sense of model quality. Table 9 shows that the valid HistGradientBoosting model improved paired absolute error over persistence by 1.51 ug/m3, with a bootstrap 95% confidence interval of 1.40 to 1.60 ug/m3. Error also increased with PM2.5 severity: the very-high regime had mean monthly MAE 29.91 ug/m3.

Table 8
Leakage stress-test diagnostic

Diagnostic	MAE	RMSE	R2	Apparent reduction %	Valid for claims
Valid chronological	15.614	29.012	0.895	0.0	True
Deliberately leaky target	0.853	5.412	0.996	94.536	False

Table 9
Statistical inference and pollution-regime error summary

Evidence	Scope	Result
Paired model comparison	HistGradientBoosting vs persistence	Mean absolute-error improvement 1.506 ug/m3; 95% CI 1.405 to 1.602; Wilcoxon p=1.40e-38
Low PM2.5 regime	Target PM2.5 <= 35 ug/m3	Mean monthly MAE 8.351 ug/m3 across 8 months and 26697 records
Very-high PM2.5 regime	Target PM2.5 > 150 ug/m3	Mean monthly MAE 29.906 ug/m3 across 8 months and 11379 records

3.9 Practical Implications for Air-Quality Research Design

The study shows that a credible PM2.5 prediction experiment needs explicit alignment between the prediction horizon, feature timing, imputation, validation split, and interpretation. A random split can be useful as a warning baseline, but it should not be the main estimate for future-

oriented claims. For ARD, the main value is the complete research-design blueprint: public data acquisition, measurement preparation, leakage-aware feature engineering, validation architecture, metrics, uncertainty intervals, figures, tables, and validation reports.

3.10 Limitations

The results are limited to one public Beijing dataset, although the evidence now covers multiple horizons, station holdout, rolling-origin evaluation, ablations, leakage diagnostics, concentration-regime stratification, and paired statistical inference. The study used established baseline models to isolate validation-design effects, not to optimize the strongest possible forecasting architecture. Predictor importance is descriptive rather than causal. Future work should test additional cities, deep learning baselines under the same leakage controls, and newer monitoring data.

4. Conclusions

This paper presents a reproducible computational research design for PM2.5 prediction across urban monitoring sites. The evidence shows that validation design, prediction horizon, feature grouping, and concentration regime all change reported accuracy and interpretation. By separating a leakage-risk random baseline from chronological, rolling-origin, station-holdout, ablation, and leakage-stress evidence, the workflow provides a stronger basis for environmental machine-learning claims. The package supports transparent rerunning of the experiment and helps authors avoid over-optimistic reporting in structured air-quality data.

Acknowledgement
The authors acknowledge the UCI Machine Learning Repository and the original data providers for making the Beijing Multi-Site Air Quality dataset available.

Conflict of Interest
The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability
The data are publicly available from the UCI Machine Learning Repository as the Beijing Multi-Site Air Quality dataset, DOI 10.24432/C5RK5G. The package includes a download script and a dataset manifest with checksum information.

Code Availability
The reproducibility package for this manuscript is archived on Zenodo at https://doi.org/10.5281/zenodo.20643454 (DOI: 10.5281/zenodo.20643454). The archive includes the code package, scripts, tests, environment files, output tables, figures, references, and validation reports needed to reproduce the results.

AI Use Disclosure
OpenAI Codex was used for code scaffolding, data-analysis support, figure/table generation, package assembly, and language-editing assistance. The human authors must verify the study design,

numerical results, interpretation, conclusions, references, and final manuscript content before submission. AI is not listed as an author.

References

[1] Chen, Song. "Beijing Multi-Site Air Quality." UCI Machine Learning Repository, 2019. https://doi.org/10.24432/C5RK5G.
[2] Zhang, Shuyi, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. "Cautionary Tales on Air-Quality Improvement in Beijing." *Proceedings of the Royal Society A* 473 (2017): 20170457. https://doi.org/10.1098/rspa.2017.0457.
[3] World Health Organization. *WHO Global Air Quality Guidelines: Particulate Matter, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Geneva: WHO, 2021. https://www.who.int/publications/i/item/9789240034228.
[4] Makhdoomi, Ahmad, Maryam Sarkhosh, and Somayyeh Ziaei. "PM2.5 Concentration Prediction Using Machine Learning Algorithms: An Approach to Virtual Monitoring Stations." *Scientific Reports* 15 (2025): 8076. https://doi.org/10.1038/s41598-025-92019-3.
[5] Zaini, Norazian, Lee Wei Ean, Ali Najah Ahmed, and Mohd Azraai Malek. "A Systematic Literature Review of Deep Learning Neural Network for Time Series Air Quality Forecasting." *Environmental Science and Pollution Research* 29 (2022): 4958-4990. https://doi.org/10.1007/s11356-021-17442-1.
[6] Gurtepe, Irde Cetinturk, Ismail Tarik Senkal, Alper Unal, Gulen Gullu, Yeser Aslanoglu, and Julian D. Marshall. "Machine Learning-Driven Regional Prediction of PM2.5 Concentration." *Environmental Modelling & Software* 192 (2025): 106586. https://doi.org/10.1016/j.envsoft.2025.106586.
[7] Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, et al. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40, no. 8 (2017): 913-929. https://doi.org/10.1111/ecog.02881.
[8] Bergmeir, Christoph, and Jose M. Benitez. "On the Use of Cross-Validation for Time Series Predictor Evaluation." *Information Sciences* 191 (2012): 192-213. https://doi.org/10.1016/j.ins.2011.12.028.
[9] Kapoor, Sayash, and Arvind Narayanan. "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science." *Patterns* 4, no. 9 (2023): 100804. https://doi.org/10.1016/j.patter.2023.100804.
[10] Kaufman, Shachar, Saharon Rosset, Claudia Perlich, and Ori Stitelman. "Leakage in Data Mining: Formulation, Detection, and Avoidance." *ACM Transactions on Knowledge Discovery from Data* 6, no. 4 (2012): 1-21. https://doi.org/10.1145/2382577.2382579.
[11] Breiman, Leo. "Random Forests." *Machine Learning* 45 (2001): 5-32. https://doi.org/10.1023/A:1010933404324.
[12] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29, no. 5 (2001): 1189-1232. https://doi.org/10.1214/aos/1013203451.
[13] Hoerl, Arthur E., and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12, no. 1 (1970): 55-67. https://doi.org/10.1080/00401706.1970.10488634.
[14] Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825-2830. https://www.jmlr.org/papers/v12/pedregosa11a.html.
[15] Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. "Array Programming with NumPy." *Nature* 585 (2020): 357-362. https://doi.org/10.1038/s41586-020-2649-2.
[16] Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17 (2020): 261-272. https://doi.org/10.1038/s41592-019-0686-2.
[17] Hunter, John D. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9, no. 3 (2007): 90-95. https://doi.org/10.1109/MCSE.2007.55.
[18] McKinney, Wes. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, 56-61, 2010. https://doi.org/10.25080/Majora-92bf1922-00a.
[19] Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed., 2022. https://christophm.github.io/interpretable-ml-book/.
[20] Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
[21] Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *Journal of Machine Learning Research* 20, no. 177 (2019): 1-81. https://jmlr.org/papers/v20/18-760.html.

[22] Efron, Bradley. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7, no. 1 (1979): 1-26. https://doi.org/10.1214/aos/1176344552.
[23] Willmott, Cort J., and Kenji Matsuura. "Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance." *Climate Research* 30 (2005): 79-82. https://doi.org/10.3354/cr030079.
[24] Hyndman, Rob J., and George Athanasopoulos. *Forecasting: Principles and Practice*. 3rd ed. OTexts, 2021. https://otexts.com/fpp3/.
[25] Gu, Jing, Bin Yang, Michael Brauer, and K. Max Zhang. "Enhancing the Evaluation and Interpretability of Data-Driven Air Quality Models." *Atmospheric Environment* 246 (2021): 118125. https://doi.org/10.1016/j.atmosenv.2020.118125.
[26] Moreno-Torres, Jose G., Troy Raeder, Rocio Alaiz-Rodriguez, Nitesh V. Chawla, and Francisco Herrera. "A Unifying View on Dataset Shift in Classification." *Pattern Recognition* 45, no. 1 (2012): 521-530. https://doi.org/10.1016/j.patcog.2011.06.019.
[27] Quionero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, eds. *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2009. https://mitpress.mit.edu/9780262170055/dataset-shift-in-machine-learning/.
[28] Tashman, Leonard J. "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review." *International Journal of Forecasting* 16, no. 4 (2000): 437-450. https://doi.org/10.1016/S0169-2070(00)00065-0.
[29] Hyndman, Rob J., and Anne B. Koehler. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting* 22, no. 4 (2006): 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.
[30] Arlot, Sylvain, and Alain Celisse. "A Survey of Cross-Validation Procedures for Model Selection." *Statistics Surveys* 4 (2010): 40-79. https://doi.org/10.1214/09-SS054.
[31] Gneiting, Tilmann, and Adrian E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102, no. 477 (2007): 359-378. https://doi.org/10.1198/016214506000001437.
[32] Zheng, Yu, Furui Liu, and Hsun-Ping Hsieh. "U-Air: When Urban Air Quality Inference Meets Big Data." In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1436-1444, 2015. https://doi.org/10.1145/2783258.2788573.