Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD38297

Urban PM2.5 Forecasting with Machine Learning: A Comparative Analysis

Zhaoxin Xu

International Leadership of Texas Garland High School, Garland, USA
jxu61942@gmail.com

Abstract. Published comparisons of PM2.5 prediction models reach different conclusions about the value of deep learning. This review examines eleven empirical studies and one methodological review published between 2018 and 2024 to assess those differences. The papers are compared by prediction target, input data, and the alternatives included in each experiment. Support vector regression and linear regression performed well in a Nottingham study using features extracted from recent concentration trends. Other studies reported lower errors with combined convolutional and recurrent networks or Transformers. In these systems, however, changes to the predictor often accompanied changes to feature selection or the representation of spatial information. The reported improvement therefore cannot always be attributed to network architecture alone. Forecast duration presents a further difficulty: estimating the next day's mean concentration and predicting an hourly peak require different evaluations. Conventional regression provides a useful baseline for these comparisons. A proposed replacement needs testing at the intended lead time, including pollution episodes, with inputs restricted to information available when the prediction is issued.

Keywords: PM2.5, machine learning, deep learning, LSTM, air quality prediction

1. Introduction

1.1. The urban forecasting task

Urban PM2.5 forecasting includes several distinct prediction tasks. An estimate for the next hour can describe the development of a pollution episode at a monitoring station, whereas a forecast of the next day's mean describes average exposure conditions. Accuracy on one target does not establish accuracy on the other, so the concentration target has to remain part of any comparison between the reported errors.

PM2.5 consists of airborne particles with aerodynamic diameters of 2.5 micrometers or less. A concentration measured within a city reflects local emissions as well as weather and transport from surrounding areas. The recent record contains information about how these conditions have affected the monitor so far. A sudden change originating elsewhere may be harder to anticipate from that record alone.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD38297

Published model comparisons use a wide range of datasets and evaluation procedures. A review of 118 deep learning studies identified this variation as a problem for assessing PM2.5 prediction methods [1]. An apparently lower error may reflect differences in the target or test data. Comparisons within the same experiment are therefore especially useful when judging the value of a model.

1.2. Reasons for comparing model families

Conventional models usually receive pollution history as a set of variables, including lagged concentrations and recent trends. Measurements from other stations can add information about conditions beyond the prediction site. In an hourly Beijing study, spatial clustering and dependence variables were incorporated into a support vector regression (SVR) model [2].

Deep networks can learn temporal patterns from sequences, but carefully constructed input features may already describe much of that pattern. A Nottingham study found that SVR and multiple linear regression outperformed the tested deep networks when given features extracted from recent PM2.5 trends [3]. This result motivates a closer look at how the models use their inputs, as well as how their architectures differ.

1.3. Research questions

Three questions organize the comparison. Which experiments favor conventional methods, and which favor deep networks (Research Question 1, RQ1)? Do these findings hold for different lead times and for hourly versus daily concentration targets (Research Question 2, RQ2)? How much of the reported difference can be interpreted in terms of the inputs, spatial coverage, and evaluation procedure (Research Question 3, RQ3)? The last question is especially relevant to hybrids, whose component methods are often evaluated together.

2. Literature review

2.1. Review approach

Eleven empirical papers from 2018-2024 form the basis of the comparison, with a methodological review used to examine evaluation practices. Their targets are numerical PM2.5 concentrations. Ten concern cities or metropolitan areas, and one covers the United States. The selection spans several model families but does not constitute a systematic search of all papers published in this period.

The papers are compared using their study locations, prediction targets, input variables, and reported errors. Particular attention is given to the alternatives tested in each experiment, since an improvement over a standalone LSTM has a different interpretation from an improvement over conventional regression. Error values from separate papers remain attached to their original datasets and forecast tasks. The analysis uses the published results without repeating the experiments or retraining the models.

2.2. Changes in forecasting methods

A Beijing experiment used a convolutional neural network (CNN) followed by a long short-term memory (LSTM) network for the next day's mean PM2.5 concentration. Its input construction used correlations across time and space, allowing the prediction to draw on more than the recent sequence at an individual station [4].

DOI: 10.54254/2755-2721/2026.LD38297

In South Korea, random forest (RF) feature importance was used to select variables before a weighted bidirectional LSTM produced forecasts one hour ahead [5]. An experiment at twelve Beijing stations instead compared an optimized Transformer with CNN-LSTM-Attention and reported better performance for the Transformer [6]. These examples differ in both predictor design and the treatment of the observations supplied to it.

2.3. Conventional prediction methods

2.3.1. Support vector regression

SVR fits a continuous response while allowing a specified tolerance around its predictions. A kernel enables nonlinear relationships to be fitted. For forecasting, lagged concentrations can be entered as separate variables, along with recent trends or measurements from nearby stations. The model then estimates the target concentration from that description of earlier conditions.

In Nottingham, fifteen attributes describing trends over the preceding twelve hours were supplied to nine conventional methods and seven deep learning methods [3]. Such inputs have already condensed the temporal record into a set of predictors. The favorable regression results apply to that representation, leaving open how the same methods would compare when supplied with less processed observations.

2.3.2. Random forests and boosting

RF combines the outputs of many decision trees. Boosting, including XGBoost, builds an ensemble in stages so that later steps improve the fit of the existing model. Both can accommodate nonlinear relationships among lagged pollution values and environmental variables. A table of such predictors is a natural input, but its construction determines which parts of the time series are available to the model.

In a comparison of nine methods across the United States, RF and SVR performed best among candidates that included XGBoost and LSTM [7]. This broader result, together with the Nottingham experiment, supports retaining conventional regressors as baselines for a proposed deep model.

RF can also be used before a recurrent network, as in the South Korean study. The forest selects the variables, while the network estimates the future concentration. Evaluating only the final network would miss part of the work done by the complete system [5].

2.4. Deep learning methods

2.4.1. LSTM and hybrid networks

An LSTM maintains a memory state as it processes a sequence of observations. Gates regulate which information is retained and how the state changes at each step. This gives the network a way to represent temporal relationships directly from the sequence. The history supplied to the network still has to be chosen: a longer window includes earlier conditions, but those conditions may have little relevance to the forecast.

A bidirectional LSTM produces forward and reverse representations of a historical sequence. Reversing the sequence changes the order of processing, not the date of the observations: both representations can be formed from measurements already available at the forecast origin. The South Korean weighted LSTM uses this architecture after RF-based feature selection [5].

DOI: 10.54254/2755-2721/2026.LD38297

CNN-LSTM passes the features extracted by convolution to a recurrent network. The positions supplied to the convolution may represent times or stations, depending on the input arrangement. The Qingdao comparison reported a root mean square error of 8.216 μg/m³ for CNN-LSTM, below the values for CNN alone (11.356) and LSTM alone (14.367). The combined model also achieved a coefficient of determination of 0.91 [8].

Preprocessing is more extensive in a Beijing model that combines empirical mode decomposition, a denoising autoencoder, and a bidirectional LSTM [9]. The decomposition represents different patterns in the concentration series, while the autoencoder is intended to reduce noise before prediction. An error reduction for this complete method measures the effect of the combined procedure. The recurrent stage's individual contribution requires a separate comparison.

2.4.2. Transformers and spatial information

A Transformer uses attention to weight relationships within its input. For PM2.5 prediction, the input may include measurements from several times and stations. This gives the model access to changes beyond a single monitoring site. The value of that wider coverage depends on whether the additional observations describe conditions relevant to the target location.

In Greater Los Angeles, a spatiotemporal Transformer combined recent pollution measurements with weather, traffic, temporal indicators, and wildfire information. The model tracked abrupt concentration changes more effectively during wildfire conditions [10]. Information about the fire may help explain this result, since it describes a pollution source that is not fully represented by the recent history at a local monitor.

MSAFormer was tested in the Haidian district of Beijing. It first encodes meteorological observations from several sites, then uses a Transformer to predict PM2.5 [11]. The encoding stage reduces a large set of related weather variables to a smaller representation. Its contribution should be considered alongside that of the prediction network when interpreting the result.

A study of Los Angeles County combined a time-series Transformer with output from the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem). A quadtree divided the county into areas of different sizes. The model used a 48-hour history to predict the following 48 hours, with variable-resolution grids performing better than the fixed grids tested [12]. Changing the grid alters how observations enter the prediction. This experiment consequently provides evidence about the spatial arrangement of the inputs as well as the forecasting model.

3. Comparative analysis

3.1. Differences in reported accuracy

The Qingdao results show an advantage for the combined network over its individual components. Calculated from the reported RMSE values, CNN-LSTM reduced error by approximately 27.7% relative to CNN and 42.8% relative to LSTM [8]. This comparison does not establish how it would perform against SVR. The Nottingham experiment, which did include conventional regression, favored SVR and multiple linear regression when the inputs described recent concentration trends [3]. For RQ1, the choice of comparison models therefore limits the conclusion that can be drawn from each result.

Input preparation offers a possible explanation for these differences. A predictor supplied with trend attributes receives temporal information in an explicit form; a network supplied with a sequence must learn a representation of that information. Comparing both input forms on one

DOI: 10.54254/2755-2721/2026.LD38297

dataset would test whether the Nottingham finding depends on feature construction. For hybrid methods, removing one stage at a time would address a different question: whether the improvement remains after convolution, denoising, or another component is removed.

3.2. Reading the error measures

Mean absolute error (MAE) is the average magnitude of the prediction errors. Root mean square error (RMSE) squares each error, averages the squared values, and then takes the square root. Large deviations therefore have a stronger effect on RMSE. Reading the two measures together can help identify whether occasional large errors contribute substantially to a model's performance. Both are expressed in μg/m³ when predictions and observations are on the original concentration scale. Errors on normalized data require conversion before comparison in these units.

For the coefficient of determination (R^2), the reference is the variation of the observations around their mean. A value of 0.91, as reported in Qingdao, corresponds to a residual sum of squares equal to 9% of that reference variation. If the residual sum exceeds the reference, R^2 becomes negative. Neither value describes a proportion of forecasts judged individually correct.

3.3. Summary of the empirical evidence

Table 1 summarizes the eleven empirical studies. The target column distinguishes an established lead time from the resolution of the observations. Labels such as 'hourly series' describe the available data and should not be read as a claim that the model predicts one hour ahead.

Table 1. Summary of representative PM2.5 forecasting studies

Study	Setting	Target or scale	Approach and inputs	Reported comparison
Yang et al. (2018) [2]	Beijing	Hourly concentration	Spatially enhanced SVR; local and spatial predictors	Spatial representation improved conventional SVR.
Pak et al. (2020) [4]	Beijing	Next-day daily mean	CNN-LSTM; pollution, weather, and spatial correlations	Joint spatial and temporal representation supported prediction.
Wood (2022) [3]	Nottingham	Current time; 1, 3, 6, 12 h ahead	Nine conventional and seven deep models; fifteen trend attributes	SVR and linear regression led the tested deep models.
Kim et al. (2023) [5]	South Korea	1 h ahead	Weighted bidirectional LSTM; RF-selected variables	Feature selection and weighting improved the tested forecasts.
Cui et al. (2023) [6]	Beijing	Hourly series	Transformer and CNN-LSTM-Attention; multi-station observations	Transformer outperformed the tested hybrid comparator.
Yu et al. (2023) [10]	Greater Los Angeles	Hourly series	Spatiotemporal Transformer; pollution, weather, traffic, fire data	Improved tracking of abrupt changes during wildfire conditions.
Teng et al. (2023) [9]	Beijing	Multi-step; includes 12 h ahead	Decomposition, denoising, and bidirectional LSTM; PM2.5 history	The combined method improved longer-term prediction.
Vignesh et al. (2023) [7]	United States	Daily concentration series	Nine methods, including boosting, RF, SVR, and LSTM	RF and SVR were strongest among the models tested.
Wang et al. (2023) [11]	Haidian, Beijing	Hourly series	MSAFormer; multi-site weather encoding and PM2.5 history	The combined representation improved tested prediction results.

DOI: 10.54254/2755-2721/2026.LD38297

Table 1. (continued)

Bai et al. (2024) [8]	Qingdao	Station concentration series	CNN-LSTM; spatial and temporal feature representation	CNN-LSTM outperformed standalone CNN and LSTM.
Zhang and Yu (2024) [12]	Los Angeles County	Following 48 h	Transformer; sensors, WRF-Chem output, quadtree grids	Variable-resolution grids outperformed tested fixed grids.

Note: Series labels indicate observation frequency where a common lead time is not established. Comparisons and errors belong to the individual studies.

Beijing accounts for five of the eleven empirical papers. The remaining studies cover Qingdao, Nottingham, South Korea, two Los Angeles areas, and the United States. This concentration of evidence in a few settings limits conclusions about how the methods would perform in other cities.

3.4. Lead time and averaging

An hourly dataset does not define a one-hour forecast. It describes the interval between observations. The input window gives the amount of history supplied, while the lead time measures how far ahead the prediction extends. Without these details, the labels "short-term" and "long-term" say little about the actual task.

The Nottingham experiment evaluated the current time and predictions one, three, six, and twelve hours ahead [3]. By contrast, the Beijing CNN-LSTM study predicted the following day's mean concentration [4]. Averaging over a day can smooth out an hourly peak, so an error for the daily mean does not describe how well the peak was predicted. Current-time estimation is a separate task again, because it provides no advance warning.

A forecast several hours ahead must estimate changes that occur before new measurements can be incorporated. Recursive methods use earlier predictions as inputs to subsequent steps, which can propagate an initial error through the forecast. Predicting several future values in one step avoids this feedback, although changes in emissions and weather still have to be estimated. Adding more historical observations does not necessarily resolve that uncertainty.

The Beijing decomposition model obtains its temporal representation from the concentration record, whereas the Los Angeles County model also receives atmospheric-model output and spatial information. Their longer forecasts therefore have different sources of predictive information. For RQ2, this makes the error at each lead time more informative than a single summary across all forecast steps. Performance at one hour should not be assumed to persist at twelve hours.

3.5. Useful inputs and reliable evaluation

Recent PM2.5 measurements are useful because concentrations often persist over time. They provide less warning of a sudden increase that has not yet reached the local monitor. Weather forecasts or observations from other locations may help anticipate such a change.

Adding more variables also creates more opportunities for missing data and timing errors. A measurement from a nearby station is only available to an operational model after it has been received. Similarly, later observed weather cannot be used as though it had been known when the forecast was issued. A model tested with those later observations answers a different question from one supplied with a weather forecast available at the time.

DOI: 10.54254/2755-2721/2026.LD38297

Preprocessing requires the same attention to timing. Interpolation between two measurements uses information that would be unavailable if the second measurement had not yet occurred. Scaling and feature selection should be fitted to training data, and decomposition should use only the history available at the forecast origin. For RQ3, tests with delayed or missing measurements would further clarify the dependence on complete inputs. These are proposed evaluation checks; no independent audit of data leakage was conducted for the selected papers.

4. Discussion

4.1. Evaluation during pollution episodes

An average error over a long test period gives limited information about a brief pollution peak. A model can underestimate that peak or predict it several hours late while making small errors during the more numerous ordinary observations. An episode-level comparison should therefore examine the predicted rise against the observed sequence, including the timing of the maximum. For advance warning, a delayed peak remains a forecasting failure even if its estimated concentration is close to the measurement.

In the Greater Los Angeles experiment, the Transformer received wildfire information in addition to pollution and other environmental variables [10]. Repeating the comparison with the wildfire variables removed would test how much those inputs contribute. A further comparison could supply them to a conventional predictor. Together, these tests would help distinguish an advantage of attention from an advantage of having information about the pollution source.

Peak magnitude and timing do not capture false alarms. A method that predicts episodes frequently may detect more of the observed peaks but also produce more warnings without a corresponding event. Reporting those cases alongside missed episodes would make the consequences of model selection clearer. The preferred balance will depend on how the warnings are used.

4.2. Computation and explanation

Computational cost cannot be ranked reliably from the results assembled here because runtime measurements are not comparable. For a working forecast service, the relevant interval runs from data receipt through preprocessing to delivery of the prediction. Training is a separate expense whose importance depends on how often the model is updated. A relatively slow training procedure may be manageable when updates are infrequent, whereas an input that routinely arrives late can limit every forecast. Hardware and memory requirements are needed to interpret either comparison.

RF feature importance can indicate which variables a fitted model relies on, while attention weights describe relationships within the network's input representation. Neither, on its own, identifies the physical cause of a pollution episode. Related predictors can substitute for each other, so the apparent importance of one may change when another is removed. Checking explanations across test periods would help distinguish stable patterns from features of a particular fitted model.

4.3. Choosing models for an urban application

Model selection should begin with a defined forecast, including the location, lead time, and data available at issue time. A persistence forecast gives an inexpensive reference by carrying the most recent observation forward. A simple regression and conventional machine learning methods can

DOI: 10.54254/2755-2721/2026.LD38297

then show how much is gained from lagged values and environmental predictors. LSTM or attention-based models can be tested against that reference using equivalent information.

The errors left by a baseline can guide the next experiment. If it misses abrupt increases, additional weather or source variables can be tested on those episodes. If its accuracy falls at an excluded station, the experiment should examine how well it transfers between locations. These tests give a specific reason for changing the model. When several alternatives perform similarly over repeated test periods, the effort required to prepare their inputs and update them becomes relevant to the choice.

4.4. Limitations and further evaluation

Five of the eleven empirical studies use Beijing data, so the apparent range of methods exceeds the geographic range of the evidence. Performance elsewhere remains uncertain. The analysis is also limited to published experiments: differences in tuning and preprocessing cannot be resolved without access to the implementations and data. A gain reported for one configuration may change when those choices are held constant.

Reproducible comparisons need the processing steps that precede prediction as well as the model code. The Dataset-Method-Experiment Standard provides a way to document the dataset and experimental choices [1]. For forecasting, that record should identify the target time, the training and test periods, and when the inputs became available. A shared dataset alone leaves considerable room for differences if studies construct their features or forecast targets in different ways.

A further test would reserve an entire city for evaluation. Errors there would reveal how well a model trained elsewhere handles a different combination of weather and emission sources. Reserving a station within a city examines a more local form of transfer. Unusual pollution periods also merit separate evaluation because the conditions encountered during them may be poorly represented in the training record.

5. Conclusion

Conventional methods remain credible candidates for urban PM2.5 forecasting. Their strong results in Nottingham and the United States comparison prevent a general conclusion that deep learning is more accurate. The gains reported for CNN-LSTM and Transformer systems are nevertheless relevant to the settings in which they were tested. Interpreting those gains requires attention to the comparison model and to any changes in the inputs or preprocessing.

A model comparison is most informative when the forecast target, available observations, and test period are shared. Longer lead times and pollution episodes require their own error analysis, since a good average result may conceal failures at the times of interest. The next step is to test whether the reported advantages persist under these conditions and at locations absent from training. Until such comparisons are available, the selected studies support several useful approaches rather than a single preferred architecture.

For an operational service, testing could begin with persistence and simple regression. If these forecasts miss sudden increases or underestimate peaks, the next experiment should examine whether weather forecasts, upwind observations, or source information reduce those errors. A deep network can then be compared with the revised baseline using the same available information. Grouping studies by forecast target, input preparation, and test conditions, as in this analysis, provides a practical basis for planning and documenting these experiments. Reproducibility also depends on access to the data and processing steps. In the absence of agreed testing protocols,

DOI: 10.54254/2755-2721/2026.LD38297

comparisons on shared datasets are easier to interpret when preprocessing decisions are reported as fully as the network design.

References

[1] Zhou, S., Wang, W., Zhu, L., Qiao, Q., & Kang, Y. (2024). Deep-learning architecture for PM2.5 concentration prediction: A review. *Environmental Science and Ecotechnology, 21*, 100400. https://doi.org/10.1016/j.ese.2024.100400

[2] Yang, W., Deng, M., Xu, F., & Wang, H. (2018). Prediction of hourly PM2.5 using a space-time support vector regression model. *Atmospheric Environment, 181*, 12-19. https://doi.org/10.1016/j.atmosenv.2018.03.015

[3] Wood, D. A. (2022). Trend decomposition aids forecasts of air particulate matter (PM2.5) assisted by machine and deep learning without recourse to exogenous data. *Atmospheric Pollution Research, 13*(3), 101352. https://doi.org/10.1016/j.apr.2022.101352

[4] Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of the Total Environment, 699*, 133561. https://doi.org/10.1016/j.scitotenv.2019.07.367

[5] Kim, B., Kim, E., Jung, S., Kim, M., Kim, J., & Kim, S. (2023). PM2.5 concentration forecasting using weighted Bi-LSTM and Random Forest feature importance-based feature selection. *Atmosphere, 14*(6), 968. https://doi.org/10.3390/atmos14060968

[6] Cui, B., Liu, M., Li, S., Jin, Z., Zeng, Y., & Lin, X. (2023). Deep learning methods for atmospheric PM2.5 prediction: A comparative study of transformer and CNN-LSTM-attention. *Atmospheric Pollution Research, 14*(9), 101833. https://doi.org/10.1016/j.apr.2023.101833

[7] Vignesh, P. P., Jiang, J. H., & Kishore, P. (2023). Predicting PM2.5 concentrations across USA using machine learning. *Earth and Space Science, 10*(10), e2023EA002911. https://doi.org/10.1029/2023EA002911

[8] Bai, X., Zhang, N., Cao, X., & Chen, W. (2024). Prediction of PM2.5 concentration based on a CNN-LSTM neural network algorithm. *PeerJ, 12*, e17811. https://doi.org/10.7717/peerj.17811

[9] Teng, M., Li, S., Yang, J., Wang, S., Fan, C., Ding, Y., Dong, J., Lin, H., & Wang, S. (2023). Long-term PM2.5 concentration prediction based on improved empirical mode decomposition and deep neural network combined with noise reduction auto-encoder - A case study in Beijing. *Journal of Cleaner Production, 428*, 139449. https://doi.org/10.1016/j.jclepro.2023.139449

[10] Yu, M., Masrur, A., & Blaszczak-Boxe, C. (2023). Predicting hourly PM2.5 concentrations in wildfire-prone areas using a SpatioTemporal Transformer model. *Science of the Total Environment, 860*, 160446. https://doi.org/10.1016/j.scitotenv.2022.160446

[11] Wang, H., Zhang, L., & Wu, R. (2023). MSAFormer: A Transformer-based model for PM2.5 prediction leveraging sparse autoencoding of multi-site meteorological features in urban areas. *Atmosphere, 14*(8), 1294. https://doi.org/10.3390/atmos14081294

[12] Zhang, S., & Yu, M. (2024). Enhanced urban PM2.5 prediction: Applying quadtree division and time-series transformer with WRF-chem. *Atmospheric Environment, 337*, 120758. https://doi.org/10.1016/j.atmosenv.2024.120758