

Physically consistent dual-attention bidirectional long short-term memory for hourly PM_{2.5} forecasting with atmospheric signal conditioning

Ibrahim Ouchen^{1*}, Soufiana Mekouar¹, Oussama Arab¹,
Roumaissae Azguet², Mohamed Mastere^{1,3}

¹ Department of Geomorphology and Geomatics, Scientific Institute, Mohammed V University in Rabat, Avenue Ibn Batouta, P.B. 703, Rabat-Agdal, Rabat, Morocco

² Institut National de Géophysique (ING), Centre National pour la Recherche Scientifique et Technique (CNRST), Rabat, Morocco

³ School of Public Management, Governance and Public Policy, College of Business & Economics, University of Johannesburg, South Africa

* Corresponding author's e-mail: ibrahim.ouchen@is.um5.ac.ma

ABSTRACT

This study aims to develop a physically consistent, dual-attention bidirectional Long Short-Term Memory (BiLSTM) framework for accurate hourly forecasting of PM_{2.5} concentrations in New Taipei City, Taiwan, and to determine whether multi-scale atmospheric signal conditioning and attention-based temporal modelling can improve forecasting performance and generalisation across heterogeneous coastal-mountain atmospheric regimes. Hourly PM_{2.5}, co-located gaseous pollutant, and meteorological observations from five monitoring stations, were pre-processed through wavelet-based denoising, Box–Cox variance stabilisation, and robust scaling. A residual dual-attention BiLSTM incorporating adaptive feature gating, dual global pooling, and a softplus output layer was trained using Huber loss and the Adam optimiser, and benchmarked against a baseline BiLSTM and traditional statistical models (VAR and ARIMA). Forecasts were combined through an inverse-error-weighted ensemble, and performance was assessed using MSE, RMSE, MAE, MAPE, R², and explained variance, complemented by an ablation analysis at each station. The proposed model achieved R² between 0.80 and 0.85 across all five stations, consistently outperforming the baseline BiLSTM and the ARIMA/VAR benchmarks, which produced near-zero or negative explanatory power. Improvements associated with individual architectural components. The model shows attenuation for extreme concentrations (>50–60 µg/m³), residual heteroscedasticity persists at high pollution levels, and the results are based on a single metropolitan area over the study period, so wider spatial and temporal validation would strengthen the generalisability of the conclusions. The framework provides a computationally efficient, physically constrained tool for hourly PM_{2.5} forecasting that can support real-time public health advisories and emission-management decisions in coastal-mountain metropolitan areas. The study introduces a station-adaptive, physically consistent dual-attention BiLSTM that combines signal conditioning, adaptive gating, and error-weighted statistical ensembling without requiring predefined spatial graph structures, offering new insight into the conditions under which architectural complexity benefits PM_{2.5} forecasting.

Keywords: PM_{2.5} forecasting, hybrid deep learning, bidirectional LSTM with dual attention, atmospheric signal conditioning, urban air quality dynamics.

INTRODUCTION

Environmental and other external stressors can have substantial implications for population health, reinforcing the importance of reliable

environmental monitoring and risk assessment (Ali et al., 2024). Fine particulate matter (PM_{2.5}) is one of the major global environmental health concerns because of its ability to penetrate deeply into the respiratory system and its association

with cardiovascular and respiratory diseases (Lelieveld et al., 2015; Pope and Dockery, 2006). Accurate short-term forecasting of hourly PM_{2.5} concentrations is important for public health advisories, real-time emission mitigation, and urban air-quality management. However, hourly PM_{2.5} forecasting is particularly challenging because pollutant concentrations are shaped by complex interactions among meteorology, atmospheric chemistry, and anthropogenic emissions. These interactions become especially pronounced in coastal-mountain metropolitan environments, where sea-land breezes, orographic flows, and monsoon-related transport can produce nonlinear and rapidly changing pollution dynamics (Seinfeld and Pandis, 2016).

New Taipei City, Taiwan, represents such a complex environment, with dense urban emissions occurring within a topographically varied region influenced by coastal and mountainous atmospheric conditions. In the present study, five monitoring stations Banqiao, Linkou, Tamsui, Wanli, and Xizhi are modelled individually to examine the reliability of the forecasting framework under different atmospheric and pollution scenarios. Rather than relying on predefined spatial graph structures, the framework evaluates whether the models can generalise across different coastal-basin transport regimes. The spatial heterogeneity among the monitoring locations therefore provides an appropriate setting for assessing whether a forecasting approach can learn the dynamic behaviour of PM_{2.5} while maintaining predictive performance across different atmospheric regimes.

Air-quality forecasting has traditionally relied on time-series models such as autoregressive integrated moving average (ARIMA) and vector autoregressive (VAR) models. These approaches can perform satisfactorily under relatively stationary and linear conditions but are less suited to representing complex nonlinear chemical processes, unexpected regime changes, and interactions among multiple meteorological conditions. Machine-learning methods, including support vector regression, random forest, and boosting techniques, have demonstrated greater ability to represent nonlinear relationships but generally have limited temporal memory for capturing long-range dependencies (Grange et al., 2018; Jiang et al., 2004). Deep-learning methods have consequently been increasingly applied to PM_{2.5} forecasting. Recurrent neural networks,

particularly long short-term memory (LSTM) models, can capture temporal dependencies and nonlinear patterns in pollution data (Hochreiter and Schmidhuber, 1997). Bidirectional LSTM (BiLSTM) models further enhance contextual representation by processing sequential information in both forward and backward directions. Attention mechanisms have subsequently been incorporated to emphasise informative temporal and spatial information and to dynamically assign greater importance to relevant drivers (Bahdanau et al., 2014; Luong et al., 2015).

More recently, transformer-based and graph-based architectures have further advanced PM_{2.5} forecasting. Sparse-attention Transformer networks have demonstrated improved modelling of long-term dependencies in complex pollutant time series (Zhang and Zhang, 2023). Hybrid deep-learning approaches combining graph neural networks (GNNs) with attention mechanisms have been proposed to capture spatial correlations among monitoring sites and have frequently achieved lower forecasting errors in multi-city settings (Guan et al., 2025). Adaptive Graph-Transformer models integrating graph convolutions, temporal convolutions, and self-attention have also been reported to achieve state-of-the-art municipal-scale PM_{2.5} forecasting with fine spatial extrapolation (Velez-Serrano and Alvaro-Meca, 2025). In addition, transport-informed GNNs have incorporated advection-diffusion physics into learnable graph operators to improve operational PM_{2.5} forecasting (Dimri et al., 2026).

Despite these advances, several limitations remain in current spatio-temporal deep-learning approaches to air-quality forecasting. Transformer- and graph-based architectures can require large sensor networks or high-resolution spatial grids to fully exploit their representational capacity, which may not be available in local urban monitoring networks. Although Transformers are effective at modelling long-range dependencies, they can be sensitive to noise and irrelevant variables in multivariate pollution data. GNNs depend on predefined graph structures, whose construction may be subjective or based on external information such as geographical proximity, and such structures may not adapt adequately to regime shifts associated with nonstationary meteorological conditions. In addition, many deep-learning models remain primarily data-driven and do not systematically address the nonstationary and multi-scale characteristics of environmental

time series. Consequently, high-frequency stochastic variability, long-term pollution patterns, and short-duration extreme haze events may not be represented equally effectively, while the absence of explicit physical constraints may result in physically implausible forecasts, such as negative concentrations.

These limitations leave an important methodological gap: how to achieve accurate and physically plausible hourly PM_{2.5} forecasts across heterogeneous atmospheric regimes using a forecasting framework that can account for multi-scale signal variability and dynamically identify informative temporal and meteorological–chemical features without requiring a predefined spatial graph. Addressing this gap requires not only improving predictive accuracy but also determining whether appropriate signal conditioning and attention-based temporal modelling can improve the robustness of forecasts across monitoring sites characterised by different pollution regimes.

Accordingly, the purpose of this study is to develop and evaluate a physically consistent dual-attention BiLSTM framework for hourly PM_{2.5} forecasting and to determine whether the integration of multi-scale signal conditioning, attention-augmented temporal modelling, adaptive feature selection, and statistical forecast information improves forecasting performance and generalisation across different coastal–mountain atmospheric regimes. The underlying hypothesis is that wavelet-based denoising, Box–Cox variance stabilisation, and robust scaling can provide a more suitable representation of the nonstationary PM_{2.5} signal by suppressing high-frequency stochastic noise, mitigating heteroscedasticity, reducing sensitivity to outliers, and preserving structural variations and extreme haze events. A second hypothesis is that a residual BiLSTM with dual attention, adaptive feature gating, and dual global pooling can more effectively capture both long-term temporal dependencies and short-duration pollution peaks while dynamically adjusting the influence of meteorological and chemical covariates. Finally, it is expected that imposing non-negativity through a softplus output and combining the deep-learning forecasts with VAR and ARIMA statistical backgrounds through inverse-error-weighted ensemble integration will improve the physical plausibility and overall reliability of the forecasts. To test these hypotheses, the proposed framework incorporates wavelet-based denoising, Box–Cox variance stabilisation,

robust scaling, a residual BiLSTM backbone, dual attention, adaptive feature gating, dual global pooling, and a softplus output layer, followed by inverse-error-weighted integration with VAR and ARIMA forecasts. The framework is evaluated independently at Banqiao, Linkou, Tamsui, Wanli, and Xizhi to determine whether the proposed approach can maintain forecasting performance across distinct coastal-basin transport and atmospheric regimes.

STUDY AREA

The area of study comprises five districts in Figure 1: Xizhi, Banqiao, Linkou, Tamsui, and Wanli. These were chosen to represent New Taipei and its metropolitan surroundings, each presenting varying environmental and human factors that contribute to predicting PM_{2.5} concentrations. For example, there are busy traffic routes, mixed residential and industrial zones, new suburban development, coastal weather conditions, and semi-rural backgrounds. Banqiao and Xizhi have considerable vehicle emissions, Linkou shows the impact of recent expansion and higher emission sources, while Tamsui, on Taiwan's northern coast, is significantly affected by sea airflow and monsoon dynamics. Wanli, though less urbanised, faces threats from regional pollution and transport in the Taipei basin. This diversity makes these districts effective for modelling PM_{2.5} changes and creating a prediction framework (Tsai et al., 2021).

METHODOLOGY

In this paper, a physically aware deep learning framework is presented for hourly PM_{2.5} concentration predictions in New Taipei City, Taiwan, a metropolitan area with high levels of urbanisation, diverse industrial emissions, vehicle pollutants, and complex coastal mountain circulation and seasonal northeast monsoon transport (Chou et al., 2023). The combination of these elements yields highly nonlinear, nonstationary, and regime-dependent pollution dynamics, providing major hurdles for classic statistical and solely data-driven models. To address these challenges, we propose a hybrid residual dual-attention Bidirectional Long Short-Term Memory (BiLSTM) architecture coupled with robust atmospheric signal

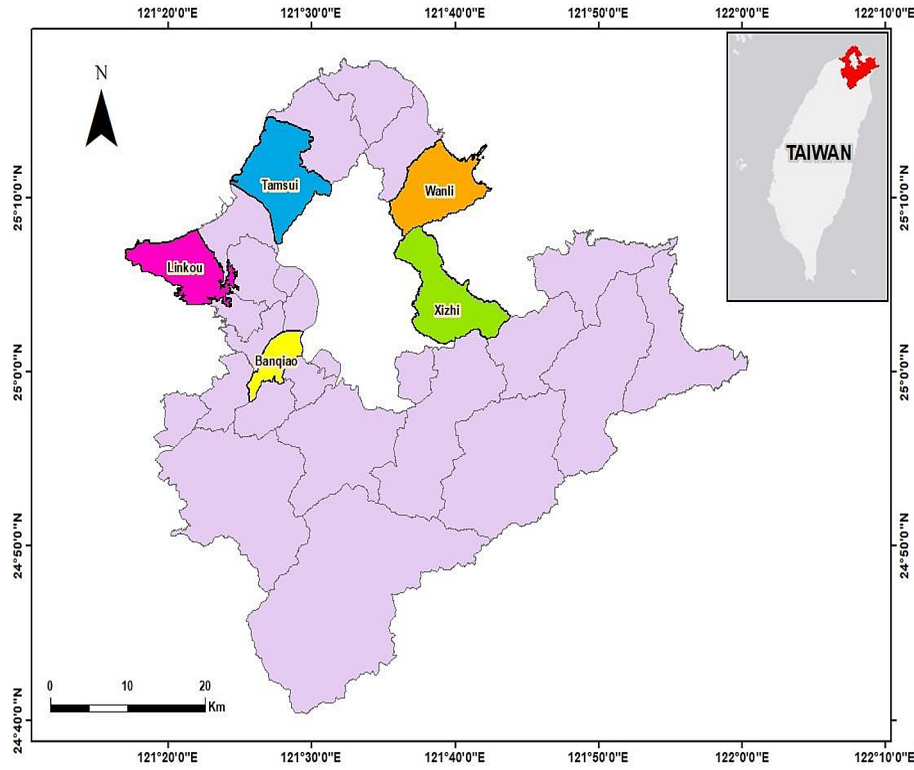


Figure 1. Study region exhibiting PM2.5 concentration forecast districts in New Taipei City, Taiwan: Banqiao, Xizhi, Linkou, Tamsui, and Wanli

conditioning and statistical ensemble benchmarking. Hourly observations were obtained from around five permanent monitoring sites. For station s and time step t , the predictor vector is:

$$X_t^{(s)} = [AQI_t^{(s)}, SO2_{2,t}^{(s)}, CO_t^{(s)}, O3_t^{(s)}, NO2_{2,t}^{(s)}, WS_t^{(s)}, WD_t^{(s)}] \quad (1)$$

and the target variable is:

$$y_t^{(s)} = PM2.5_t^{(s)} \quad (2)$$

The forecasting objective is:

$$\hat{y}_{t+1}^{(s)} = f(X_{t-L+1:t}^{(s)}), L = 24 \quad (3)$$

Atmospheric signal conditioning

Prior to being employed in a model, raw PM2.5 observations pass through a hierarchical signal conditioning procedure that separates random fluctuations, different types of variance, and measurement errors. First, we use discrete wavelet decomposition (Gao and Yan, 2010) to separate multi-scale temporal features from high-frequency observational noise:

$$x_t^{(l)} = W^{-1}(T \times {}_\lambda(W(x_t^{raw}))) \quad (4)$$

where: W denotes the discrete wavelet transform and $T \times \lambda$ is a soft-thresholding operator to attenuate spurious transients without discarding physically meaningful fluctuations.

For the purpose of stabilising non-constant variance and accommodating the nonlinear distributional characteristics of urban particulate concentrations, a Box–Cox transformation (Sakia, 1992) is used with an additive positivity shift. Thus, to maintain physical interpretability and output plausibility, model training is conducted on Min–Max scaled original PM2.5 values, despite the fact that this step enhances statistical homoscedasticity.

Median imputation is applied to handle missing values, and robust scaling is performed to mitigate the influence of outliers for predictor variables, which include meteorological covariates and gaseous pollutants.

$$x_t^{scaled} = (x_t - median(x)) / IQR(x) \quad (5)$$

ensuring resilience to extreme events, episodic outliers, and regime-dependent skewness inherent in urban air quality data.

For the VAR model (Lütkepohl, 2013), in particular, principal component analysis (PCA) (Greenacre et al., 2022) is conducted on the multivariate predictors to achieve dimensionality reduction while retaining principal components that explain 95% of total variance. This ensures that the dominant co-variational patterns in atmospheric drivers are captured while retaining computational tractability for multivariate autoregressive modelling.

Finally, chronological splitting is conducted to split the data into training and testing sets, with 70% for training and 30% for testing to prevent information leakages. The 24-hour sliding windows are then constructed to capture diurnal cycles, pollutant accumulations, and transient transport phenomena in sequences that are conducive to bidirectional modelling. This ensures that the proposed learning framework comes with inputs that are physically accurate and temporally coherent, thus achieving a balance between interpretive accuracy and statistical rigour.

Residual dual-attention BiLSTM architecture

The nonlinear combination of emission accumulation, turbulence dispersion, chemical transformation, and airborne transport mechanisms is responsible for urban PM_{2.5} dynamics. To capture these multi-scale, regime-dependent interactions, we combine a series of multivariate atmospheric states into a high-dimensional latent space using a bidirectional LSTM encoder:

$$H_t = [\vec{h}_t; \hat{h}_t] \quad (6)$$

where: $\vec{h}_t$ and $\hat{h}_t$ denote causal accumulation and anticipatory contextual dependence, respectively. In this latent space, forward propagation is used to represent the sequential accumulation of pollutants based on the effects of previous emissions and meteorological forcing conditions, and backward propagation allows the model to learn anticipatory corrections based on the subsequent temporal structure, thus forming a bidirectional dynamical system.

Adaptive feature gating is represented as a multiplicative state-dependent operator:

$$g_t = \sigma(W_g H_t + b_g), \tilde{H}_t = H_t \cdot g_t \quad (7)$$

It performs an input-conditioned rescaling of latent features, selectively amplifying latent features associated with chemically dominating

processes and suppressing irrelevant features during stable dispersion regimes. This gating mechanism applies an input-conditioned, nonlinear transformation conditioned on latent state, modulating the trajectory in latent space via feature-wise reweighting.

The attention, following the Luong-style dot product definition, computes a context vector:

$$c_t = \sum_i \alpha_{t,i} h_i \quad (8)$$

interpreted as an attention-weighted aggregation over the latent trajectory that emphasizes historically relevant states as determined by learned attention weights. By integrating attention-derived context with the original latent space via residual fusion:

$$H'_t = H_t + c_t \quad (9)$$

The architecture retains dynamical patterns while adding a contextually corrective perturbation that stabilizes gradient flow and mitigates information decay over deep temporal layers.

Dual global pooling applies the enhanced residual latent sequence onto a compact representation:

$$z = [\text{AvgPool}(H'), \text{MaxPool}(H')] \quad (10)$$

performing a multi-scale aggregation that simultaneously records low-frequency baseline dynamics and high-amplitude episodic excursions.

The last prediction layer applies a Softplus activation:

$$\hat{y}_t = \ln(1 + e^{W_{oz}z + b}) \quad (11)$$

It maintains non-negativity while retaining differentiable, numerically stable gradients. Thus, this residual dual-attention BiLSTM can be interpreted as a hybrid operator-based dynamical system: it maps multivariate atmospheric trajectories to a physically informed latent space, performs regime-adaptive reweighting and temporal integration, and produces output-level projections that combine deep learning flexibility with physical plausibility.

The architecture of our proposed approach is illustrated in Figure 2. It represents the proposed framework as a structured transformation in which multivariate atmospheric inputs are progressively mapped into a constrained scalar forecast through sequential representation. The diagram depicts a hierarchical state-space transformation. The Bidirectional LSTM produces a latent dynamical embedding of the

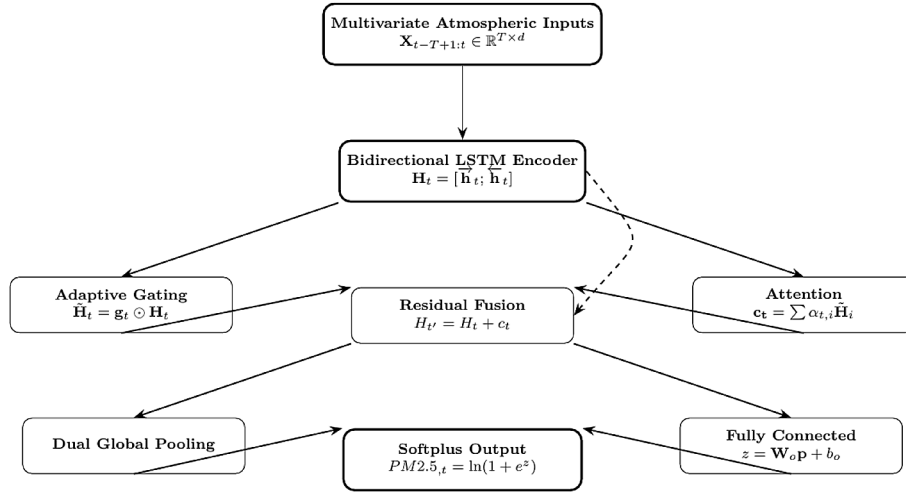


Figure 2. Residual dual-attention BiLSTM framework for hourly PM_{2.5} forecasting

24-hour atmospheric track, providing a temporally indexed hidden latent space that captures persistence and delayed nonlinear interactions. The adaptive gating functions as a multiplicative state-dependent operator, focusing on this latent space and reweighting latent coordinates in response to the development of regime parameters. The attention mechanism then performs a learnt projection over the temporal axis, yielding a context vector that approximates a data-driven temporal significance measure. The residual preserves representational continuity by confining attention-induced updates relative to the original encoded state, hence stabilising gradient propagation and reducing distortion of long-memory structures. Dual global pooling then carries out functional reductions across the temporal dimension, with mean aggregation approximating integrated background loading, while max aggregation maintains extreme-state excursions, yielding a compact yet regime-sensitive representation. Finally, the Softplus output layer enforces positivity, incorporating physical feasibility directly inside the optimisation landscape.

Model optimization and ensemble

We perform parameter estimation by minimising the Huber loss function, a piecewise objective that combines the sensitivity of the squared error for small variations with the resilience of the absolute error for large residuals.

$$L_{\delta}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2, \text{ for } |y - \hat{y}| \leq \delta;$$

$$L_{\delta}(y, \hat{y}) = \delta(|y - \hat{y}| - \frac{1}{2}\delta), \text{ for } |y - \hat{y}| > \delta \quad (12)$$

This technique reduces the disproportionate influence of infrequent severe pollution occurrences while being sensitive to regular diurnal oscillations, ensuring that the learned mapping encompasses both routine and episodic dynamics.

The Adam method is used for optimisation, along with gradient clipping and a learning rate decrease. Gradient clipping reduces unstable updates in deep recurrent models, especially in bidirectional LSTM encoders, where temporal dependencies may spread exponentially over extended sequences. Adaptive learning-rate scheduling enhances convergence in the initial epochs while allowing for smaller gradient steps during subsequent refinement, resulting in numerically stable and physically plausible representations of pollutant evolution. The proposed optimised BiLSTM, baseline BiLSTM, VAR, and ARIMA predictions have been combined into an inverse-error weighted ensemble to make the final forecasts more reliable. The mean absolute error (MAE) is used to figure out the ensemble weighting.

$$w_m = (1/MAE_m) / \sum_j (1/MAE_j), \hat{y}_{ensemble}$$

$$= \sum_m w_m \hat{y}^{(m)} \quad (13)$$

providing more weight to models that show better validation accuracy. This method uses the strengths of each model: the BiLSTM models capture nonlinear and temporally lagged pollutant interactions, the VAR model captures dominant multivariate co-variational modes, and the ARIMA model offers a simple autoregressive approach as a baseline for trend extrapolation. The ensemble combines the biases and variances of the models, producing physically plausible, consistent

and statistically robust hourly forecasts of PM_{2.5}, which are useful in the management and formulation of policies related to air quality in cities.

EXPERIMENTS

Air quality observations were derived from the data collected at five air quality monitoring stations in New Taipei City (Xizhi, Wanli, Banqiao, Tam-sui, and Linkou), which are representative of urban environments with coastal-mountain circulation and heterogeneous human emissions. The data set contains concentrations of PM_{2.5} and co-located gaseous species (SO₂, CO, NO₂, and O₃) and meteorological parameters (wind speed and wind direction), measured on an hourly basis. These observations represent the combined response of emission activity, atmospheric chemical transformation, and meteorological transport processes controlling the variability of particulate air quality.

To maintain the temporal causality of the data, the data set was split chronologically into 70% for training and 30% for testing. The PM_{2.5} time series was denoised using a discrete wavelet transform, and variance stabilisation was applied via a Box-Cox transformation with a positive shift. For the vector autoregressive baseline model, the predictor variables were median imputed and scaled to minimise the effect of outliers. For the vector autoregressive baseline model, the predictor variables were also subject to a PCA transformation to retain 95% of the variance.

The prediction of the proposed approach is based on a residual dual-attention bidirectional long short-term memory network. It is based on recognising nonlinear temporal relationships, which are part of the nature of pollutants. The inputs are taken on a 24-hour cycle. It includes an adaptable feature that allows for changing the nature of the prediction. Temporal attention is incorporated to identify previous periods. It is more stable to include a residual relationship. Dual global pooling is effective for background pollution levels and random peaks. The addition of a Softplus layer is necessary to keep the predictions between a certain range of non-negative concentration values.

Optimising parameters for the model involved using the Adam optimiser. This involved Huber loss, gradient clipping, and a learning rate schedule. An ablation study was performed to assess the contributions of attention, residual fusion, feature gating, and pooling components.

Classical statistical baselines, including vector autoregression (VAR), autoregressive integrated moving average (ARIMA), and a standard BiLSTM architecture, were trained using identical sequences. Final forecasts were integrated through an inverse-error weighted ensemble.

Model performance was evaluated using mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), coefficient of determination (R²), and explained variance.

Reproducibility statement. To support independent verification, all preprocessing and modelling steps described above (chronological 70/30 train-test split, wavelet denoising, Box-Cox stabilisation, robust scaling, and PCA for the VAR baseline) were applied identically and deterministically at each of the five stations. The residual dual-attention BiLSTM, baseline BiLSTM, VAR, and ARIMA models were trained independently per station using fixed architecture and optimisation settings. The proposed BiLSTM architecture, there were 128 bidirectional LSTM cells, recurrent dropout rate of 0.1, dropout rate of 0.25, L₂regularisation parameter of 1×10^{-4} , and a 64 cell dense layer with dropout of 0.2. Training of the model was done via Adam optimizer with gradient clipping (clipnorm = 0.5), Huber loss, batch size of 128, up to 150 epochs. Early stopping mechanism was implemented for 8 epochs patience period according to the validation loss. The learning rate was set to 0.001 for epoch 1–20, 0.0005 for epoch 21–50, 0.0002 for epoch 51–80, and 0.0001 afterward. Look-back window length was equal to 24. The baseline BiLSTM architecture comprised two LSTM layers bidirectionally, each having 100 units, followed by dropout layer with probability of 0.4. The training process was done using MSE as loss function and Adam as optimiser algorithm with learning rate of 0.001, batch size 32, and maximum epochs set to 100. The pre-processing pipeline used db4 wavelet noise removal at level 3, RobustScaler for feature scaling, and PCA. The experiments were done by using the defined experiment setup instead of using multiple seed-based experiments. Thus, no unsupported number of runs or mean $\pm$ standard deviation from multiple runs is used. The ablation tests between different architectures were done using the same chronological split for testing at each station. The performance of models was evaluated based on MSE, RMSE, MAE, MAPE, R², and EVS. The comparison presented here

corresponds to the performance of models on the held-out chronological test dataset and hence can be taken as a descriptive comparison of their predictive performance rather than as proof of statistical significance of superior performance across multiple training runs. The processed station-level datasets as well as full code for the modeling process consisting of preprocessing, model training and validation, ablation studies, and testing can be made available by the corresponding author upon request.

Results

In Figure 3a, we can see the observed and predicted concentration levels at the Wanli station. Most of the points have excellent alignment along the 1:1 reference line. A slight compression of the points at high levels of concentration is due to the attenuation of severe pollution levels. This is normal in nonlinear sequence models that use powerful loss functions.

In Figure 3b, we have plotted the comparison of residuals and the anticipated concentration levels at the Wanli station. The spread of the residuals is increasing as we move to higher levels of concentration. Therefore, we can say that there is some level of heteroscedasticity. The residuals have a natural tendency to remain negative during peak levels of concentration.

Figure 4a displays the comparison between the measured and predicted values at the Tamsui station. Most of the data points are distributed around the 1:1 line, suggesting a good overall

calibration ($R^2 = 0.828$). The model is consistently performing well in the range of low to moderate concentration levels ($5\text{--}45\ \mu\text{g m}^{-3}$). At higher concentration levels ($> 60\ \mu\text{g m}^{-3}$), a sign of saturation is observed in the figure, where the model predictions are slightly underestimating the extreme peak values.

Figure 4b displays the plot of the residuals vs. the predicted concentration values. The residuals are centred around zero, suggesting a lack of bias in the model predictions. However, the spread in the residuals increases at higher concentration levels, indicating a possible heteroscedasticity issue in the data. The more negative values in the residuals are visible at the peak levels, which is consistent with the slight underestimation of the peak.

Figure 5a compares observed and predicted concentrations at the Linkou station. The points cluster closely around the 1:1 reference line, indicating good overall agreement ($R^2 = 0.801$). Model performance is strongest within the main concentration range ($5\text{--}40\ \mu\text{g m}^{-3}$). At higher concentrations ($> 60\ \mu\text{g m}^{-3}$), a mild compression appears, suggesting partial attenuation of extreme pollution episodes while maintaining overall proportionality.

Figure 5b shows residuals versus predicted concentrations at Linkou. We observe that the majority of the residuals cluster around zero. Therefore, no standard bias is seen. The heteroscedasticity is observed when the range increases with the quantity. Moreover, negative residuals are detected more frequently when peak occurrences are experienced, indicating that peak values are overestimated.

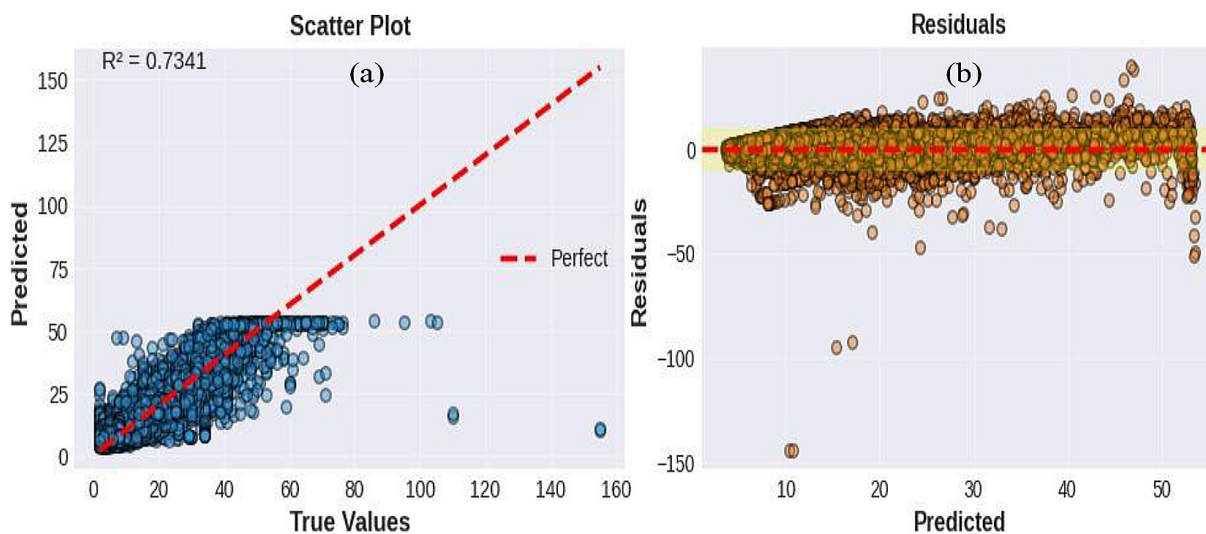


Figure 3. Evaluation of the optimized PM2.5 prediction model for Wanli (a) observed versus predicted pollutant concentrations illustrating model calibration; (b) residuals vs. predicted concentrations

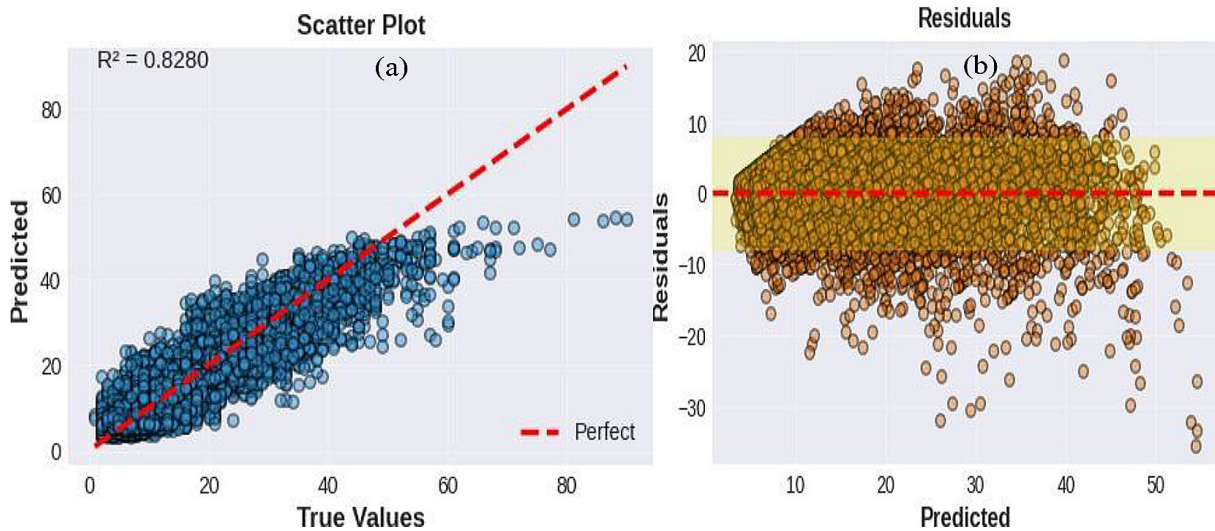


Figure 4. Evaluation of the optimized PM2.5 prediction model for Tamsui: (a) observed versus predicted pollutant concentrations; (b) residuals vs. predicted concentrations

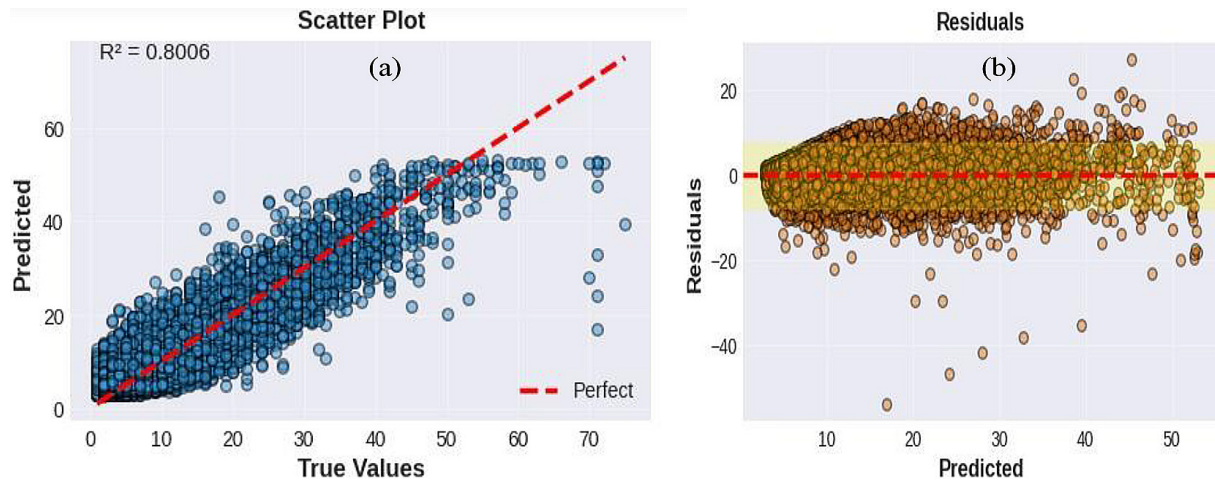


Figure 5. Evaluation of the optimized PM2.5 prediction model for Linkou: (a) observed versus predicted pollutant concentrations; (b) residuals vs. predicted concentrations

In Figure 6a, we observe a plot of observed and anticipated values for PM2.5 concentrations in the Banqiao station. The dots are closely grouped along the 1:1 line, indicating a high level of predictive accuracy ($R^2 = 0.82$). The model tracked observed values well across all moderate concentrations of pollution. At higher concentrations of pollution ($> 40 \mu\text{g m}^{-3}$), we observe a slight tendency for the model to fall below the 1:1 line, showing a slight tendency to underestimate extreme pollution events.

In Figure 6b, we can see a plot of residuals vs. expected values. The residuals track along zero for decreasing values, indicating a lack of bias in predictions for normal events. As we move to higher values, we can see a significant increase in

scatter, indicating a high level of heteroscedasticity in events involving higher levels of pollution. The residuals seem to be symmetric, indicating a lack of significant nonlinear effects.

As shown in Figure 7a, observed and anticipated values of PM2.5 concentrations are plotted for the Xizhi station. The model has a very good predictive ability ($R^2 = 0.854$), as demonstrated by a very tight linear relationship between observed and anticipated values, especially for a range of 0 to $40 \mu\text{g m}^{-3}$. At higher concentrations ($> 50\text{--}60 \mu\text{g m}^{-3}$), a flattening of the values is observed, whereas observed values are increasing.

Figure 7b presents a plot of residuals vs. expected values. It is observed that residuals tend to be centred around zero for lower concentrations,

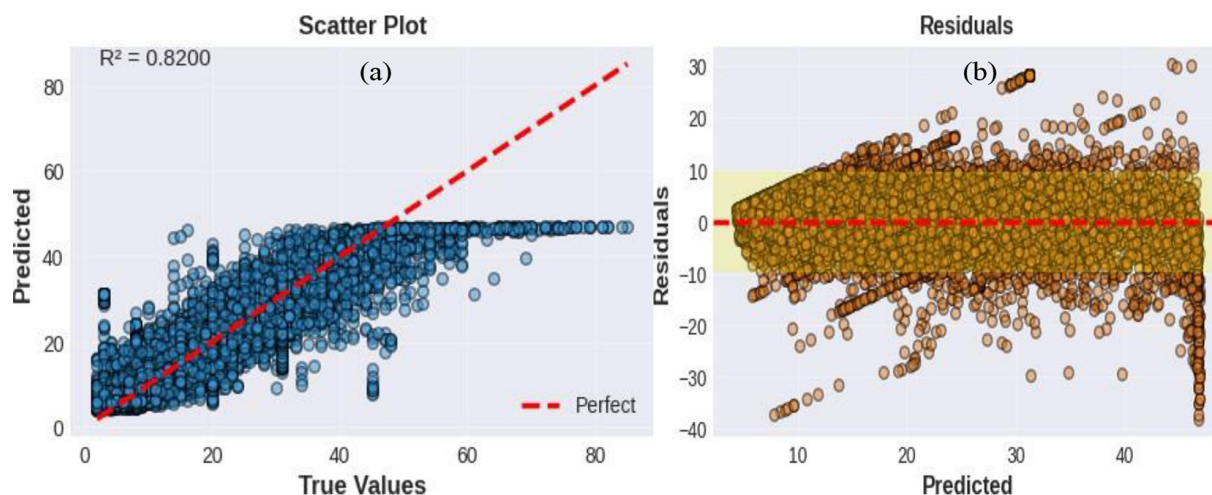


Figure 6. Evaluation of the optimized PM2.5 prediction model for Banqiao: (a) observed versus predicted pollutant concentrations; (b) residuals vs. predicted concentrations

indicating unbiased predictions for normal circumstances. At higher concentrations, a significant increase in residuals is observed, indicating a high level of heteroscedasticity. There is a lack of curvature in residuals, indicating that the major nonlinear interactions are correctly addressed in the model.

Ablation study

The ablation study across all five stations shows that while architectural components enhance performance, this is dependent on heterogeneous atmospheric regimes rather than monotonically increasing performance.

For Tamsui (Table 1), adding temporal attention improves performance as measured by R^2 ($0.8215 \rightarrow 0.8319$), showing that temporally localised dependencies are important in this coastal

city. However, performance decreases with the full architecture, showing that additional context modules may be over-parametrising smooth pollution dynamics driven by marine environments.

At Linkou (Table 2), the full architecture provides the best improvement ($R^2 = 0.7941$), with reductions in both RMSE and MAPE. The weaker attention-only configuration implies that temporal weighting alone is insufficient, and that residual fusion and nonlinear feature gating capture more complex accumulation processes.

A similar pattern appears at Xizhi (Table 3), where the full configuration improves performance ($R^2 = 0.8668$). Thus, the combined mechanisms of residual fusion and dual pooling appear to better represent both background accumulation and episodic peaks, reflecting multi-scale pollution dynamics.

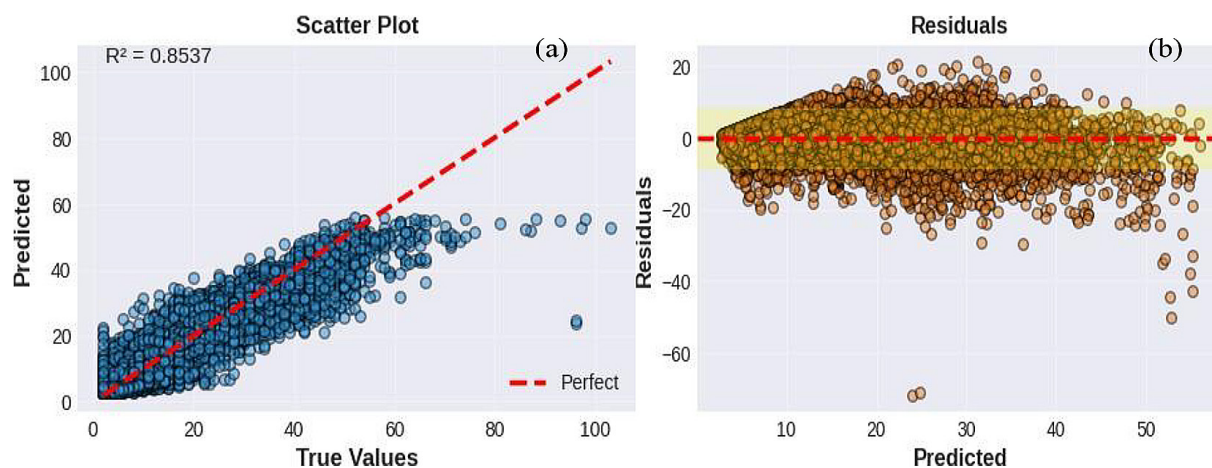


Figure 7. Evaluation of the optimized PM2.5 prediction model for Xizhi: (a) observed versus predicted pollutant concentrations; (b) residuals vs. predicted concentrations

Table 1. Ablation results at Tamsui station

Model	MSE	MAE	RMSE	R ²	MAPE (%)
Baseline	15.5280	2.8423	3.9406	0.8215	29.3243
+ Attention	14.6231	2.7929	3.8240	0.8319	29.0155
Full Model	16.0131	2.8802	4.0016	0.8160	30.6871

Table 2. Ablation results at Linkou station

Model	MSE	MAE	RMSE	R ²	MAPE (%)
Baseline	17.4974	3.0629	4.1830	0.7667	45.2810
+ Attention	18.4289	3.1268	4.2929	0.7543	46.2364
Full Model	15.4465	2.9209	3.9302	0.7941	39.5248

Table 3. Ablation results at Xizhi station

Model	MSE	MAE	RMSE	R ²	MAPE (%)
Baseline	17.2943	2.8820	4.1586	0.8510	30.1587
+ Attention	17.4298	2.9126	4.1749	0.8498	30.5169
Full Model	15.4552	2.7822	3.9313	0.8668	28.7384

Table 4. Ablation results at Banqiao station

Model	MSE	MAE	RMSE	R ²	MAPE (%)
Baseline	22.5107	3.2613	4.7445	0.8190	31.5806
+ Attention	20.8582	3.1460	4.5671	0.8323	30.3564
Full Model	22.6288	3.2358	4.7570	0.8181	30.2319

In contrast, Banqiao (Table 4) reveals improvement with attention but no additional benefit from the whole design, demonstrating that in emission-dominated urban contexts, further nonlinear modulation provides only slight improvement.

Wanli (Table 5) exhibits little sensitivity to architectural complexity, with near-equal results across configurations. This demonstrates that pollution fluctuation at this coastal–mountain interface is mostly regulated by wider advective processes rather than complicated local nonlinear interactions.

Figure 8 shows the comparison of the absolute error distributions of the five models at Xizhi station. The proposed optimised model has the most compact error distribution in terms of absolute errors, where the median is close to zero,

and the interquartile range (IQR) is small. The Baseline and ensemble models have similar performance to the optimised model. The ARIMA model demonstrates larger dispersion in the error distribution. The VAR model has much worse performance in terms of the median and range of the error distribution.

Figure 9 shows a similar ranking at Wanli, however, with a greater total error range reflecting the station's broader PM_{2.5} fluctuation. The optimised model has the lowest median error with the smallest IQR, while the baseline and ensemble models are similar with a slightly larger dispersion. ARIMA has the largest error dispersion, and VAR maintains the worst performance with the largest median error and unpredictability. The

Table 5. Ablation results at Wanli station

Model	MSE	MAE	RMSE	R ²	MAPE (%)
Baseline	25.1210	3.2731	5.0121	0.7254	26.9197
+ Attention	24.9045	3.2795	4.9904	0.7278	27.3808
Full Model	25.1281	3.3522	5.0128	0.7254	27.8820

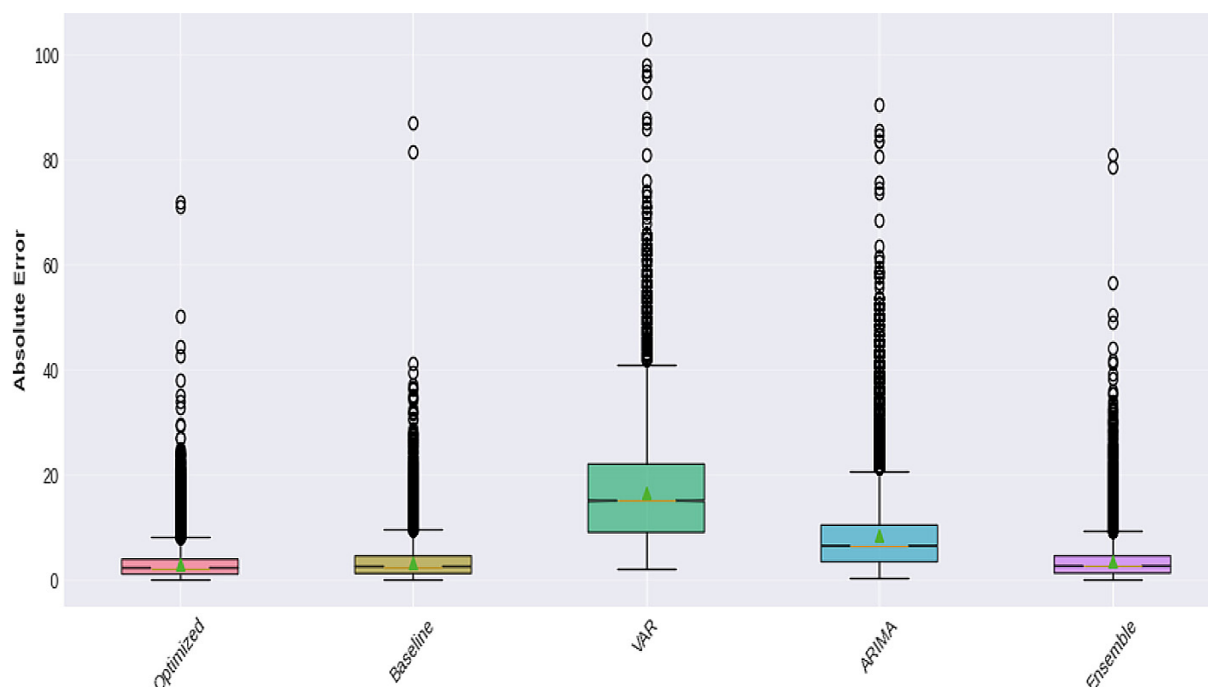


Figure 8. Distribution of absolute prediction errors for five forecasting models (Optimized, Baseline, VAR, ARIMA, Ensemble) at Xizhi applied to PM2.5 data

large gaps between the models refer to the problem of anticipating extreme events.

At Tamsui (Figure 10), the PM2.5 concentration has relatively small variations. The optimised model maintains the most compact distribution, while the baseline and ensemble models perform similarly well. ARIMA produces much larger

errors, although VAR likewise has the highest median error and dispersion. Occasional large outliers correspond to episodic pollution events.

As shown in Figure 11, we can observe that in Linkou, the optimised model has the smallest median error and IQR range, while baseline and ensemble remain competitive, with much higher

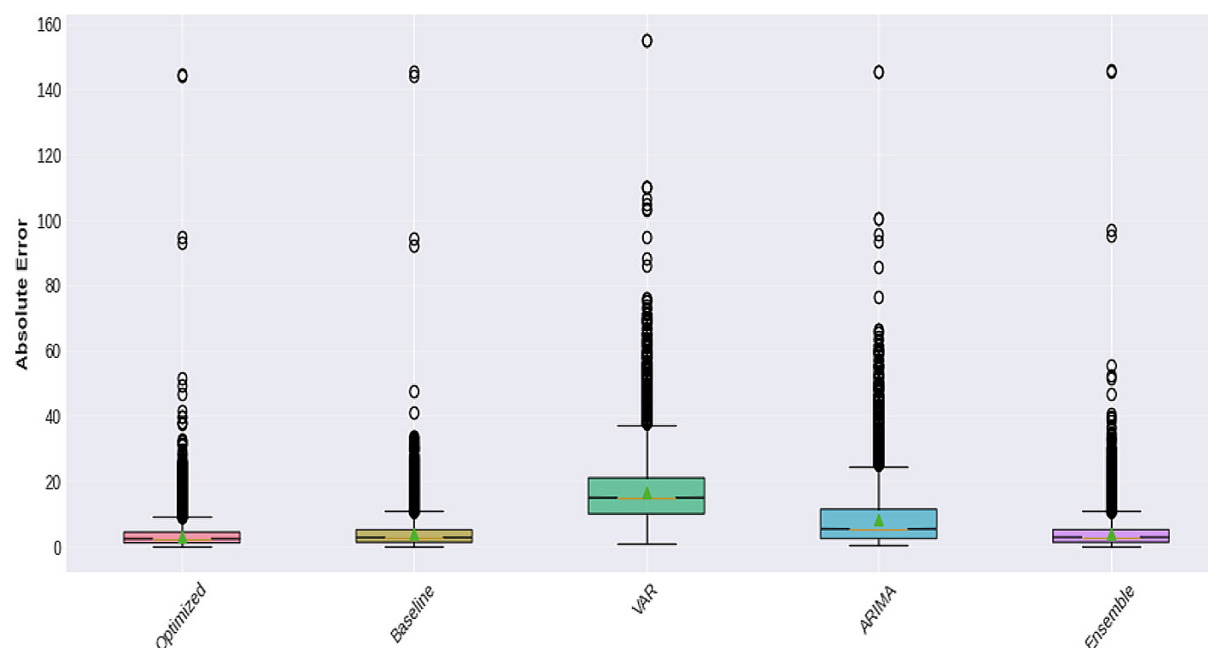


Figure 9. Absolute error distributions for five forecasting models (Optimized, Baseline, VAR, ARIMA, Ensemble) evaluated on PM2.5 at Wanli

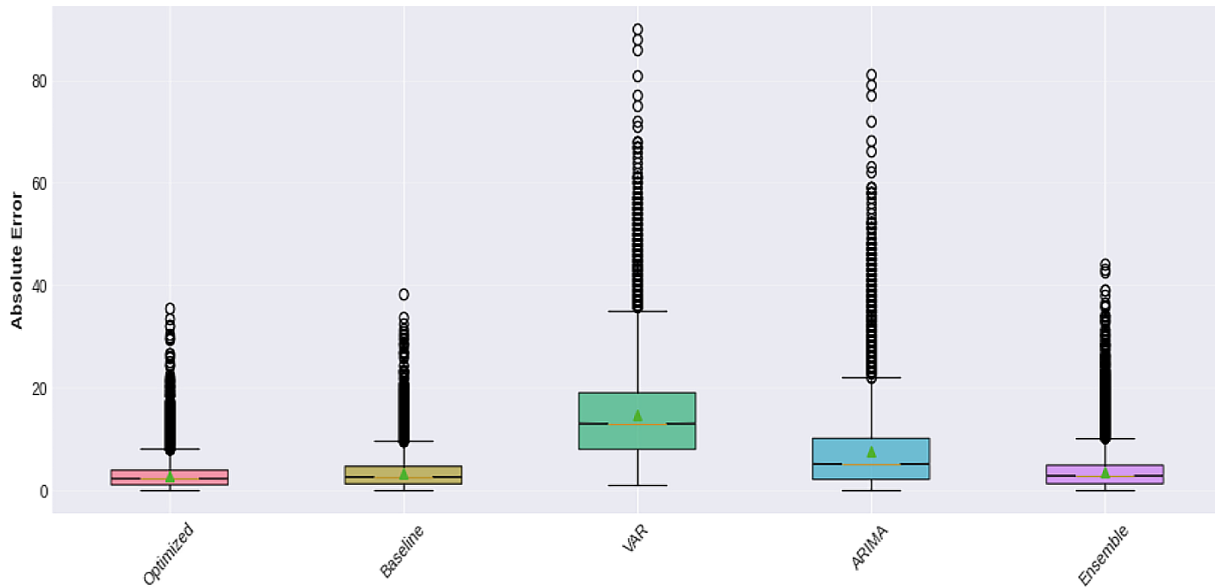


Figure 10. Absolute error distributions for five forecasting models (Optimized, Baseline, VAR, ARIMA, Ensemble) evaluated on PM2.5 at Tamsui

dispersion. ARIMA has higher variation, while VAR always has the largest error. Outliers in all models suggest that there are episodic spikes in concentration that can affect prediction accuracy.

In Banqiao (Figure 12), which has the widest range of errors of all stations, the optimised model shows the best stable performance with low median error and small IQR. The baseline and ensemble models remain close competitors but show much larger variability. The ARIMA model is in the middle position, although the VAR model shows poor performance with the largest median errors and outliers.

Model comparison and ensemble behavior across stations

The predictive performance and associated ensemble weights are shown in Tables 6–10. Throughout all monitoring locations, the proposed optimised residual dual-attention BiLSTM consistently achieves the best performance.

For the Wanli station, the optimised model's performance is better than that of traditional linear models, which have negative explanatory power. The latter's negative explanatory power implies that they have failed to encapsulate the problem's

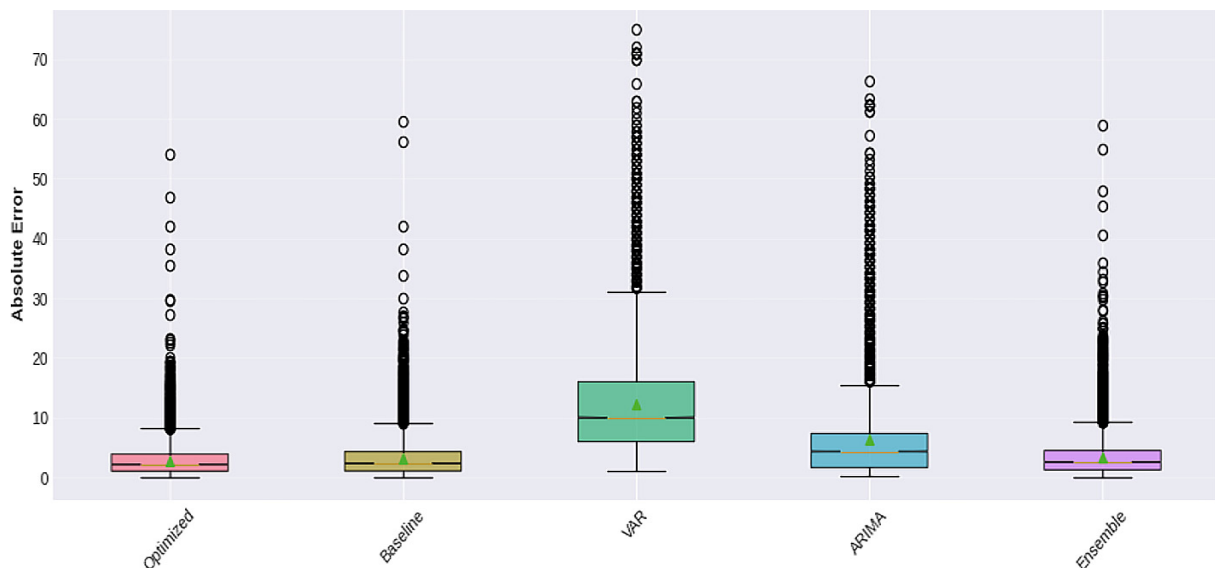


Figure 11. Absolute error distributions for five forecasting models (Optimized, Baseline, VAR, ARIMA, Ensemble) evaluated on PM2.5 at Linkou

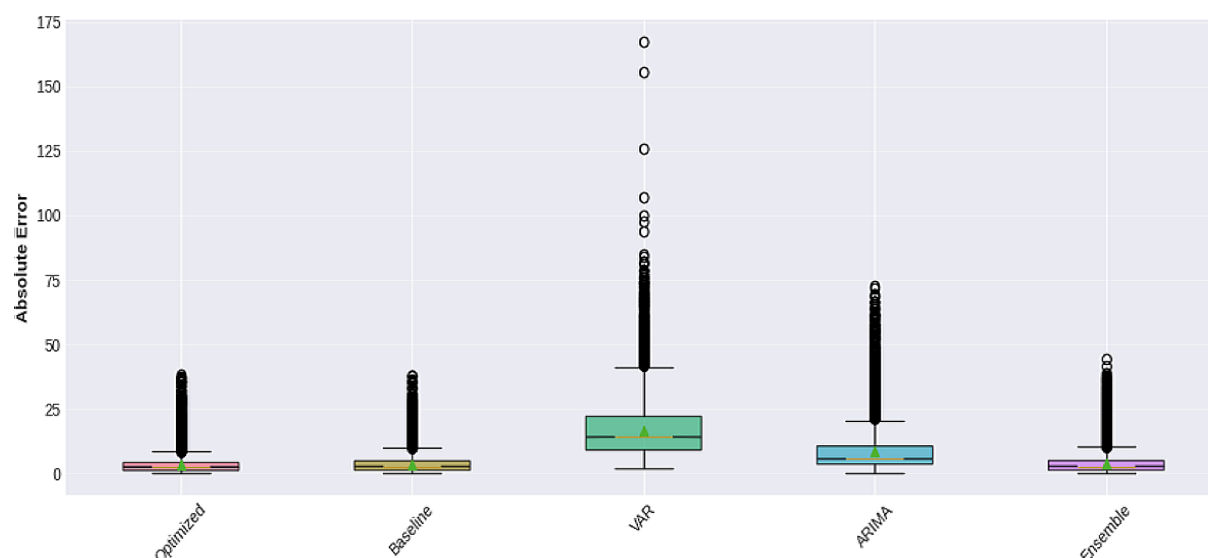


Figure 12. Absolute error distributions for five forecasting models (Optimized, Baseline, VAR, ARIMA, Ensemble) evaluated on PM2.5 at Banqiao, the highest-volatility site

Table 6. Model performance at Wanli station

Model	MSE	MAE	RMSE	R ²	MAPE (%)	EVS
Optimized (Proposed)	24.3271	3.3071	4.9323	0.7341	27.8935	0.7354
Baseline	29.4699	3.7826	5.4286	0.6779	30.6891	0.6791
Ensemble	32.1223	3.8773	5.6677	0.6489	27.4814	0.7133
ARIMA	138.5395	8.1920	11.7703	-0.5141	47.6467	0.0000
VAR	363.0079	16.4307	19.0528	-2.9674	100.1214	-0.0347

underlying dynamics. The performance of the ensemble model is no better than that of the optimised model. Thus, it can be concluded that in the context of advection-dominated coastal environments, the addition of linear models only contributes unnecessary variability. The high weight of the improved model in the ensemble model (0.4033) is another reflection of this difference in predictive capabilities.

A similar trend is seen in Tamsui, where the divergence between nonlinear sequence modelling and linear autoregressive methods is notably accentuated. The improved model enhances R² by almost 7% relative to the baseline. Furthermore, convergence warnings in ARIMA highlight the instability of purely parametric approaches in non-stationary marine pollution scenarios. Similar to Wanli, the ensemble exhibits worse performance compared to the optimised model, suggesting that hybridisation fails to mitigate the structural constraints of linear approaches.

At Linkou, where pollutant dynamics are impacted by recirculation, the optimised architecture again produces large benefits over both

baseline and linear models. Although MAPE remains relatively high, the increase in R² implies a greater capacity to capture nonlinear persistence. The ensemble provides no further advantage and modestly decreases performance, showing that the deep architecture alone appropriately depicts the prevailing system behaviour.

Xizhi has the highest explanatory power among all stations, with R² = 0.8537. Thus, the optimized model outperforms all alternatives, revealing that the dual-attention residual approach captures both multi-scale temporal relationships and episodic pollution spikes efficiently. In contrast, linear models exhibit negative explanatory ability, underscoring the importance of nonlinear representations.

At Banqiao, characterised by dense urban emissions, the improved model produces consistent improvements over the baseline via better nonlinear temporal encoding. However, consistent with observations at Wanli and Tamsui, the ensemble does not surpass the performance of the optimised model, demonstrating little complementarity between nonlinear and linear

Table 7. Model performance at Tamsui station

Model	MSE	MAE	RMSE	R ²	MAPE (%)	EVS
Optimized (Proposed)	14.9637	2.8522	3.8683	0.8280	28.3561	0.8295
Baseline	21.1217	3.4028	4.5958	0.7572	36.1237	0.7574
Ensemble	25.1575	3.6226	5.0157	0.7109	30.8884	0.7669
ARIMA	121.9291	7.7018	11.0421	-0.4013	54.5461	0.0001
VAR	305.6095	14.7850	17.4817	-2.5124	100.0028	-0.0013

Table 8. Model performance at Linkou station

Model	MSE	MAE	RMSE	R ²	MAPE (%)	EVS
Optimized (Proposed)	14.9587	2.8490	3.8677	0.8006	38.4368	0.8008
Baseline	19.1142	3.1621	4.3720	0.7452	43.8626	0.7456
Ensemble	21.9414	3.3804	4.6842	0.7075	39.8642	0.7464
ARIMA	87.4430	6.2807	9.3511	-0.1657	68.3713	-0.0001
VAR	223.2463	12.1751	14.9414	-1.9761	100.0000	0.0000

Table 9. Model performance at Xizhi station

Model	MSE	MAE	RMSE	R ²	MAPE (%)	EVS
Optimized (Proposed)	16.9788	2.9130	4.1205	0.8537	28.1855	0.8538
Baseline	21.9871	3.3281	4.6890	0.8105	33.2644	0.8108
Ensemble	25.7870	3.5009	5.0781	0.7778	31.4621	0.8100
ARIMA	133.5254	8.4049	11.5553	-0.1506	77.4044	-0.0000
VAR	390.1453	16.5558	19.7521	-2.3619	99.9944	-0.0000

Table 10. Model performance at Banqiao station

Model	MSE	MAE	RMSE	R ²	MAPE (%)	EVS
Optimized (Proposed)	22.3850	3.1873	4.7313	0.8200	32.0718	0.8210
Baseline	26.1849	3.5867	5.1171	0.7895	32.6224	0.7901
Ensemble	32.4826	3.8413	5.6993	0.7389	31.8455	0.7677
ARIMA	142.1242	8.4192	11.9216	-0.1425	67.0177	0.0001
VAR	409.5907	16.7499	20.2383	-2.2927	100.2744	-0.0790

predictions. There are three key findings, obtained across all the stations. Firstly, the optimised residual dual-attention BiLSTM model has consistently achieved superior performance compared to the baseline model, which confirms the effectiveness of the proposed approach. Secondly, the linear methods, including ARIMA and VAR, have produced almost zero or negative explanatory ability, suggesting the inherent non-linear nature of the PM_{2.5} dynamics. Thirdly, the performance of the inverse error ensemble has failed to exceed

the performance of the optimal model, suggesting the non-linear model has achieved sufficient performance, and the addition of other models has resulted in lower performance.

CONCLUSION

This study confirms that the proposed residual dual-attention BiLSTM achieves the stated objective: physically consistent and accurate

hourly PM_{2.5} forecasts across five monitoring stations characterised by heterogeneous coastal-mountain atmospheric regimes ($R^2 = 0.80\text{--}0.85$), consistently outperforming the baseline BiLSTM and the linear VAR/ARIMA benchmarks. The first hypothesis, that wavelet-based denoising, Box–Cox stabilisation, and robust scaling provide a more suitable representation of the nonstationary PM_{2.5} signal, is supported by the compact, near-unbiased residual distributions obtained across all stations. The second hypothesis, that dual attention, adaptive feature gating, and dual global pooling jointly improve the capture of both long-term dependencies and short-duration pollution peaks, is only partially supported: the ablation results show that the benefit of the full architecture is regime-dependent, providing the largest gains at terrain-influenced sites such as Linkou and Xizhi, but offering little or no improvement at more stable or advection-dominated sites such as Wanli and Tamsui. The hypothesis that inverse-error-weighted ensembling with VAR and ARIMA would improve reliability was not supported, since the linear components' negative explanatory power meant that ensembling did not exceed the performance of the optimised model alone at any station.

The main novel contribution of this work is a station-adaptive analysis showing that architectural complexity should be matched to the underlying atmospheric regime rather than applied uniformly, together with the demonstration that a physically constrained residual dual-attention BiLSTM can achieve strong generalisation across heterogeneous coastal-mountain sites without requiring predefined spatial graph structures. In doing so, the study fills a gap in the existing literature between purely data-driven deep-learning forecasters and rigid statistical or graph-based approaches, by combining multi-scale signal conditioning with physically informed, regime-aware temporal modelling.

Future work should extend the framework to additional cities and longer observation periods to test spatial and temporal generalisability, incorporate uncertainty quantification to better characterise the attenuation observed at extreme concentrations, and explore light-weight graph- or transformer-based extensions where sensor density permits, so as to further improve forecasting reliability for public-health decision-making.

Founding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Ali, N., Kausar, A., Vambol, S., Kozub, P., Kozub, S., Hrynzovskyi, A. (2024). Environmental retrieval during COVID-19 lockdown in the megacity Karachi, Pakistan. *Ecological Questions*, 35(3), 1–16. <https://doi.org/10.12775/EQ.2024.032>
2. Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv*. <https://arxiv.org/abs/1409.0473>
3. Bhaskaru, O., Lalitha, K. (2026). A review of ensemble deep learning models by transfer learning for proactive and precise heart disease findings. *Biomedical Signal Processing and Control*, 117, 109592.
4. Chou, C. C.-K., Lung, S.-C. C., Hsiao, T.-C., Lee, C.-T. (2023). Regional and urban air quality in East Asia: Taiwan. In *Handbook of air quality and climate change* (pp. 1–38). Springer.
5. Dimri, R., Choi, Y., Singh, D., Park, J., Khorshidian, N. (2026). TransNet: A transport-informed graph neural network for forecasting PM_{2.5} concentrations across South Korea. *npj Clean Air*, 2(1), 12.
6. Gao, R. X., Yan, R. (2010). Discrete wavelet transform. In *Wavelets: Theory and applications for manufacturing* (pp. 49–68). Springer.
7. Grange, S. K., Carslaw, D. C., Lewis, A. C., Boletti, E., Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM₁₀ trend analysis. *Atmospheric Chemistry and Physics*, 18(9), 6223–6239.
8. Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2(1), 100.
9. Guan, X., Mo, X., Li, H. (2025). A novel spatio-temporal graph convolutional network with attention mechanism for PM_{2.5} concentration prediction. *Machine Learning and Knowledge Extraction*, 7(3), 88.
10. Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
11. Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., Shao, D. (2004). Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment*, 38(40), 7055–7064.

12. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569), 367–371.
13. Li, X., Cheng, T., Shi, S., Guo, H., Wu, Y., Lei, M., Zuo, X., Wang, W., Han, Z. (2021). Evaluating the impacts of burning biomass on PM_{2.5} regional transport under various emission conditions. *Science of the Total Environment*, 793, 148481.
14. Luong, M.-T., Pham, H., Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1412–1421).
15. Lütkepohl, H. (2013). Vector autoregressive models. In *Handbook of research methods and applications in empirical macroeconomics* (pp. 139–164). Edward Elgar Publishing.
16. Pope, C. A., III, Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, 56(6), 709–742.
17. Qin, D., Yu, J., Zou, G., Yong, R., Zhao, Q. (2019). A novel combined prediction scheme based on CNN and LSTM for urban PM_{2.5} concentration. *IEEE Access*, 7, 20050–20059.
18. Saha, R., Saif, S., Khan, T., Alabdultif, A. (2026). A critical systematic review of ensemble learning methods for predictive healthcare. *Discover Computing*, 29(1), 117.
19. Sakia, R. M. (1992). The Box-Cox transformation technique: A review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(2), 169–178.
20. Seinfeld, J. H., Pandis, S. N. (2016). *Atmospheric chemistry and physics: From air pollution to climate change*. John Wiley & Sons.
21. Tan, D. Y., Heal, M. R., Stevenson, D. S., Reis, S., Vieno, M., Nemitz, E. (2026). Long-range impacts of biomass burning on PM_{2.5}: A case study of the UK with a globally nested model. *Atmospheric Chemistry and Physics*, 26(6), 3973–3993.
22. Tsai, I.-C., Lee, C.-Y., Lung, S.-C. C., Su, C.-W. (2021). Characterization of the vehicle emissions in the greater Taipei area through vision-based traffic analysis system and its impacts on urban air quality. *Science of the Total Environment*, 782, 146571.
23. Velez-Serrano, D., Alvaro-Meca, A. (2025). Adaptive graph neural network–transformer model for high-resolution PM_{2.5} forecasting and spatial extrapolation. *Environmental Modelling & Software*, 106771.
24. Zhang, Z., Zhang, S., Zhao, X., Chen, L., Yao, J. (2022). Temporal difference-based graph transformer networks for air quality PM_{2.5} prediction: A case study in China. *Frontiers in Environmental Science*, 10, 924986.
25. Zhang, Z., Zhang, S. (2023). Modeling air quality PM_{2.5} forecasting using deep sparse attention-based transformer networks. *International Journal of Environmental Science and Technology*, 20(12), 13535–13550.