**TELKOMNIKA Telecommunication Computing Electronics and Control**
Vol. 24, No. 6, December 2026, pp. 2023~2036
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v24i6.28079 ❒ 2023

# Decomposition-aware recurrent fusion for particulate matter forecasting

**Wahyu Hardyanto[1], Siti Wahyuni[1], Aji Purwinarko[2], Isa Akhlis[1], Mahargjo Hapsoro Adi[3]**
[1]Department of Physics, Universitas Negeri Semarang, Semarang, Indonesia
[2]Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia
[3]Universitas Negeri Semarang, Semarang, Indonesia

**Article Info**

***Article history:***

Received May 30, 2026
Revised Aug 13, 2026
Accepted Aug 28, 2026

***Keywords:***

Air quality prediction
Decomposition-aware fusion
Long short-term memory
Multi-kernel moving average
PM2.5 forecasting
Time-series forecasting

**ABSTRACT**

Accurate fine particulate matter (PM2.5) forecasting supports air-quality warnings and short-term environmental decisions. This study proposes multi-kernel decomposition-aware fusion long short-term memory (MK-DAF-LSTM), a compact recurrent forecasting model. It combines a base LSTM branch with a correction branch that separates trend and residual components using moving-average kernels of 3, 5, and 7-time steps. A learned selector fuses these components, which are encoded and added to the base prediction through a learnable correction scale. Experiments used the Beijing PM2.5 single-site dataset at 1-, 3-, and 5-hour horizons and four stations from the Beijing multi-site dataset for robustness evaluation. Against LSTM, gated recurrent unit (GRU), bidirectional LSTM (BiLSTM), convolutional neural network–LSTM (CNN-LSTM), DAF-LSTM, and attention-LSTM, MK-DAF-LSTM achieved the lowest root mean square error (RMSE) and mean absolute percentage error (MAPE) and the highest coefficient of determination ($R^2$) at the 3-hour horizon (RMSE 37.4176; $R^2$ 0.8409). Across three random seeds, RMSE was 37.5933 ± 0.1971 and $R^2$ was 0.8394 ± 0.0017. Ablation experiments supported learned multi-kernel fusion over single-kernel and equal-weight alternatives. The method provides station-wise forecasting without explicitly modeling spatial dependencies.

*This is an open access article under the CC BY-SA license.*

***Corresponding Author:***

Aji Purwinarko
Department of Computer Science, Universitas Negeri Semarang
Semarang, Indonesia
Email: aji.purwinarko@mail.unnes.ac.id

## 1. INTRODUCTION

Fine particulate matter, usually denoted as PM2.5, is one of the main air pollutants monitored in urban air-quality systems because its concentration can change within a short period and can affect public health, visibility, and daily environmental decisions. Short-term forecasting is therefore needed not only to describe the current air condition, but also to support early warning and operational monitoring. Previous studies have shown that PM2.5 forecasting remains difficult when the model has to represent temporal variation and spatial heterogeneity at the same time [1]. Short-term PM2.5 forecasting has also been discussed in a practical multi-step setting for supporting proactive environmental regulation [2]. Deep-learning and conventional time-series approaches have been compared for cleaner-air prediction, showing that forecasting performance depends not only on model complexity but also on the forecasting horizon and data pattern [3]. Their study is useful here because it reminds us of that PM2.5 forecasting is not only a matter of choosing a more complex model, but also of selecting a model that fits the forecasting horizon and

data pattern. In a recent comparison, Lee *et al.* [4] also showed that statistical and long short-term memory (LSTM)-based models can behave differently across forecasting settings. These studies suggest that PM2.5 forecasting should be treated as a practical and non-trivial time-series problem rather than as a routine sequence prediction task.

Deep learning methods have been widely used for this task because air-quality observations contain non-linear and time-dependent patterns. Li *et al.* [5] used an attention-based convolutional neural network-LSTM (CNN-LSTM) model for urban PM2.5 forecasting, showing that local feature extraction, recurrent learning, and temporal weighting can be combined in one prediction pipeline. Wei *et al.* [6] developed a CNN-LSTM-attention model optimized by the whale optimization algorithm for Beijing PM2.5 prediction, which confirms the relevance of hybrid recurrent models in this domain. Kumar *et al.* [7] used gated recurrent unit (GRU), LSTM, and attention in one forecasting framework. This design shows that recurrent models can still be strengthened when the sequence representation is enriched before the final prediction stage. Aggarwal and Toshniwal [8] also reported that a hybrid deep learning framework can be useful for urban air-quality forecasting. Zhang *et al.* [9] further showed that hybrid deep learning remains relevant for PM2.5 air-quality forecasting. These works show that recurrent models remain relevant, but they also imply that the input sequence often needs to be represented more carefully before the final prediction is made.

A direct recurrent encoder has to learn rapid fluctuations and smoother temporal tendencies from the same raw sequence. This is not always ideal for PM2.5 data, where short-lived changes and trend-like behavior may appear together. Zaini *et al.* [10] showed that decomposition can improve PM2.5 forecasting by exposing temporal components before deep learning is applied. Zheng *et al.* [11] used wavelet-packet transform to form a multi-resolution representation before LSTM-based prediction. Li *et al.* [12] combined variational mode decomposition with deep neural networks for PM2.5 forecasting. Their result supports the idea that the original PM2.5 sequence may become easier to model after it is separated into more meaningful temporal components. Sreenivasulu and Rayalu [13] introduced seasonal-trend decomposition using Loess (STL)-based decomposition before CNN-bidirectional LSTM (CNN-BiLSTM)-attention prediction, which also supports the idea that PM2.5 signals should not always be processed only in their raw form.

The same motivation appears in multi-scale and multi-kernel forecasting studies. Dong *et al.* [14] demonstrated that extracting multi-scale features can improve deep learning-based PM2.5 forecasting. Xu *et al.* [15] used a multi-kernel and decomposition-based strategy for PM2.5 prediction, which is close to the motivation of this study because different kernels can expose different temporal views. Guan *et al.* [16] proposed a decomposition-based multi-step framework using hybrid deep learning, further showing that decomposition remains useful when the forecasting horizon increases. Pak *et al.* [17] also emphasized that suitable spatiotemporal intervals can influence deep learning-based PM2.5 forecasting. These studies motivate a compact model that can construct several temporal views from the same input window before recurrent encoding.

Although decomposition, attention, and recurrent learning have each been used in air-quality forecasting, several gaps remain from an implementation and modeling perspective. Decomposition-based models often depend on a single decomposition setting or use decomposed components as direct replacements for the original sequence. Attention-based recurrent models can emphasize informative time steps, but they do not explicitly separate smoother tendency from short-term fluctuation before recurrent correction. Multi-kernel designs can expose multiple temporal scales, but the usefulness of each scale may differ across input conditions and forecasting horizons. These observations motivate a conservative correction architecture in which multi-kernel trend-residual views are learned as a bounded correction to a base recurrent forecast rather than as a complete replacement for the raw-sequence predictor.

Another line of research improves air-quality prediction by explicitly modeling spatial relations, multi-source information, or multimodal data. Wang *et al.* [18] introduced a multi-site PM2.5 forecasting approach based on spatiotemporal fusion of meteorological factors. Liu and Li [19] combined graph convolution with LSTM so that relationships between monitoring stations could be represented during forecasting. Chen *et al.* [20] used a hybrid directed graph neural network with spatiotemporal fusion for regional PM2.5 prediction. Wu *et al.* [21] extended this direction by using spatiotemporal graph attention with multiple data sources. Abbasi *et al.* [22] showed that a lightweight graph framework can still exploit clustered monitoring networks and pollutant dependency. Hu *et al.* [23] and Zhang *et al.* [24] also illustrate that stronger spatial or spatiotemporal designs usually require additional mechanisms beyond a single station-wise recurrent model. These studies clarify the boundary of the present work. multi-kernel decomposition-aware fusion LSTM (MK-DAF-LSTM) does not claim to learn explicit spatial dependency. The multi-site experiment in this paper is used only as a station-wise robustness evaluation.

Based on the above discussion, this study proposes a MK-DAF-LSTM model for short-term PM2.5 forecasting. The model keeps LSTM as the recurrent backbone, but it does not rely only on a direct raw-sequence forecast. It combines a base recurrent branch with a decomposition-aware correction branch. The

correction branch forms several trend-residual temporal views using different moving-average kernels, learns the relative importance of these views, and adds a conservatively scaled correction to the base forecast. The methodological novelty of this study is therefore not the isolated use of LSTM, moving-average decomposition, attention, or gating. Rather, the contribution lies in organizing these components into a learnable multi-kernel recurrent correction architecture that preserves the base forecast while allowing trend-residual information to adjust it. In this way, the proposed model is positioned between simple recurrent forecasting and more complex spatiotemporal or multimodal systems. It aims to improve temporal representation while keeping the model suitable for an applied short-term forecasting setting.

The main contributions of this paper are summarized as follows; (i) a learnable multi-kernel decomposition-aware recurrent correction model is proposed for short-term PM2.5 forecasting. The model combines a base LSTM forecast with trend-residual correction features derived from moving-average kernels of 3, 5, and 7-time steps; (ii) the proposed architecture is evaluated on the Beijing PM2.5 single-site dataset using 1-, 3-, and 5-hour forecasting horizons. The comparison includes recurrent, convolutional-recurrent, attention-based, and decomposition-aware baselines; and (iii) the evaluation is strengthened through repeat-seed testing, ablation analysis, paired uncertainty analysis, computational-efficiency reporting, and station-wise robustness evaluation on selected Beijing multi-site air quality stations. These analyses are used to examine both predictive performance and the contribution of the proposed multi-kernel correction branch.

The rest of this paper is organized as follows: section 2 describes the proposed MK-DAF-LSTM model, datasets, experimental setting, and evaluation metrics. Section 3 presents the main forecasting results, repeat-seed stability, ablation study, station-wise robustness evaluation, and limitations. Section 4 presents the conclusion and possible directions for future work.

## 2. METHOD

### 2.1. Model intuition

The proposed model is designed to provide a richer temporal representation than a standard LSTM network. Instead of sending the raw historical sequence directly to the recurrent encoder, MK-DAF-LSTM first constructs several decomposed temporal views using different moving-average kernels. A smaller kernel is expected to preserve local fluctuations, whereas a larger kernel captures smoother temporal behavior. This design follows the practical lesson from decomposition-based PM2.5 forecasting. Zaini *et al.* [10] showed that decomposition can help deep models handle complex PM2.5 patterns. Dong *et al.* [14] emphasized the value of multi-scale feature extraction. Xu *et al.* [15] further showed that multi-kernel decomposition can be useful for PM2.5 prediction.

The main idea is therefore not to replace LSTM, but to improve the temporal representation received by the recurrent predictor. LSTM-based and hybrid recurrent models remain widely used for air-quality forecasting because they can model temporal dependency and non-linear relationships in sequential observations. Li *et al.* [5] used CNN-LSTM with attention for urban PM2.5 prediction, while Kumar *et al.* [7] combined GRU, LSTM, and attention for enhanced air-quality forecasting. MK-DAF-LSTM follows this recurrent modeling direction, but it adds a decomposition-aware correction path so that local fluctuations and smoother tendencies do not have to be learned only from the raw input sequence.

### 2.2. Problem formulation

Given a historical multivariate sequence $X \in \mathbb{R}^{B\times L\times C}$, where $B$, $L$, and $C$ denote the batch size, look-back length, and number of variables, respectively, the forecasting problem is formulated as the estimation of the PM2.5 concentration at a future horizon $h$:

$$\hat{Y}_{t+h} = F_\theta(X_{t-L+1:t}) \tag{1}$$

where $F_\theta(\cdot)$ represents the forecasting function parameterized by $\theta$. In this study, the model uses the previous 24 hours of observations to predict PM2.5 concentrations at 1-, 3-, and 5-hour horizons.

For each prediction horizon, the time series is converted into supervised input-target pairs by sliding a fixed-length historical window along the chronologically ordered observations. The input consists of the preceding $L$ time steps, whereas the target corresponds to the PM2.5 value observed $h$ hours ahead. This setting reflects a short-term multi-horizon forecasting task. Gul *et al.* [2] reported that PM2.5 forecasting becomes more difficult as the prediction horizon moves further from the observed period. Lee *et al.* [4] also found that the relative performance of forecasting models may change across experimental settings. For this reason, the present study reports the results separately for the 1-, 3-, and 5-hour horizons instead of relying on a single aggregated score.

### 2.3. MK-DAF

MK-DAF-LSTM uses two complementary forecasting paths. The first path is a base recurrent branch that produces a direct forecast from the raw input sequence. The second path is a decomposition-aware correction branch that forms trend and residual views using multiple moving-average kernels. In the implementation used in this study, the kernel set was fixed as $\mathcal{K} = \{3,5,7\}$, which represents short, medium, and smoother temporal views within the 24-hour input window. The final forecast is obtained by adding a conservatively scaled correction to the base prediction.

For the $m$-th moving-average kernel $k_m$, the trend component is computed as

$$T^{(m)} = \psi_{k_m}(X),\ k_m \in \{3,5,7\} \tag{2}$$

where $\psi_{k_m}(\cdot)$ denotes a one-dimensional moving-average operation with replicate padding. The corresponding residual component is defined as (3):

$$R^{(m)} = X - T^{(m)}. \tag{3}$$

Instead of assigning equal importance to all kernels, MK-DAF-LSTM estimates learnable kernel weights from the latest observation in the input window. The kernel weights are computed as (4):

$$\lambda = \mathrm{softmax}\big(W_{k,2}\mathrm{ReLU}\big(W_{k,1}x_t + b_{k,1}\big) + b_{k,2}\big), \tag{4}$$

where $\lambda_m$ denotes the weight of the $m$-th moving-average kernel and $x_t$ is the latest observation in the input sequence. The mixed trend and residual representations are then obtained as (5):

$$T_{\mathrm{mix}} = \sum_{m=1}^{M} \lambda_m\, T^{(m)},\ R_{\mathrm{mix}} = \sum_{m=1}^{M} \lambda_m\, R^{(m)}. \tag{5}$$

The trend representation is projected using a dense transformation, whereas the residual representation is encoded using a one-layer LSTM followed by temporal attention. Given the hidden states $H = [h_1, h_2, \cdots, h_L]$ from the residual LSTM, the attention score and normalized attention weight are computed as:

$$e_i = v_a^{\top}\tanh(W_a h_i + b_a), \tag{6}$$

$$a_i = \frac{\exp(e_i)}{\sum_{j=1}^{L}\exp(e_j)}, \tag{7}$$

and the residual context vector is:

$$z_R = \sum_{i=1}^{L} a_i\, h_i \tag{8}$$

The trend feature $z_T$ is obtained from the projected latest trend-mix representation. The fusion gate is estimated from the concatenated trend and residual features:

$$g = \sigma\big(W_{g,2}\mathrm{ReLU}\big(W_{g,1}[z_T; z_R] + b_{g,1}\big) + b_{g,2}\big) \tag{9}$$

The correction feature follows the implemented fusion order:

$$z_C = g \odot z_R + (1 - g) \odot z_T, \tag{10}$$

where $\odot$ denotes element-wise multiplication. The correction head produces $\Delta\hat{Y}$ from $z_C$. The correction scale is implemented as a learnable conservative parameter,

$$\alpha = \sigma(\beta), \tag{11}$$

where $\beta$ is initialized to -3.0. This gives an initial correction weight of approximately 0.0474, allowing the model to start from a base recurrent forecast and gradually learn the correction strength during training. The final forecast is

$$\hat{Y}_{t+h} = \hat{Y}_{\mathrm{base}} + \alpha\Delta\hat{Y}. \tag{12}$$

This formulation differs from a direct recurrent model because the LSTM-based predictor is supported by learnable multi-kernel trend-residual temporal views before the final correction is added. It also differs from a single-kernel decomposition-aware model because the relative importance of the 3-, 5-, and 7-step moving-average views is learned from the input sequence rather than fixed manually or averaged equally. The design is consistent with decomposition and multi-scale PM2.5 forecasting studies. Zheng *et al.* [11] used wavelet-packet representation before deep prediction, Li *et al.* [12] used variational mode decomposition before deep neural networks, and Guan *et al.* [16] used a decomposition-based hybrid framework for multi-step forecasting. MK-DAF-LSTM keeps the same general motivation but implements it as a compact moving-average multi-kernel recurrent correction branch.

### 2.4. Architecture block

Figure 1 illustrates the architecture of MK-DAF-LSTM. The model consists of two complementary branches. The first is a base recurrent branch that produces a direct forecast from the raw input sequence. The second is a correction branch that derives multi-kernel trend and residual representations, refines them through trend projection, residual recurrent encoding, and temporal attention, and then produces a correction term. The final PM2.5 forecast is obtained by combining the base prediction with the weighted correction.

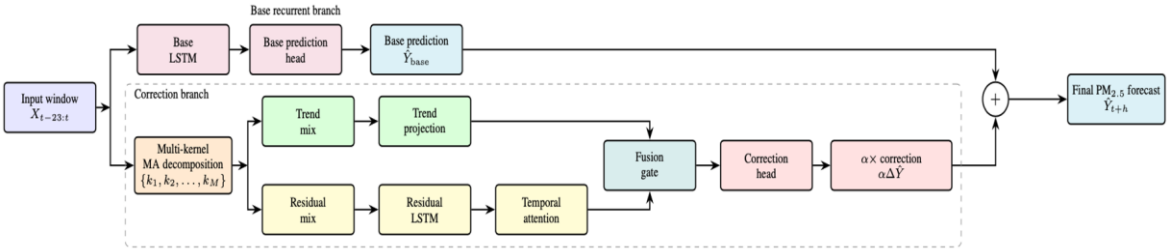



Figure 1. Architecture of MK-DAF-LSTM for PM2.5 forecasting

### 2.5. Training and forecasting procedure

Algorithm 1 summarizes the training and forecasting procedure. The algorithm begins with chronological data preparation and missing-value handling. The time series is then transformed into supervised input-target pairs according to the selected forecasting horizon. During training, each mini-batch is transformed into several kernel-based temporal representations, fused, encoded using LSTM, and optimized using the mean squared error loss.

**Algorithm 1. Training and testing workflow of MK-DAF-LSTM**

**Input:** Hourly $PM_{2.5}$ dataset $\mathcal{D}$ arranged by time; look-back length $L$; forecasting horizons $\mathcal{H} = \{1,3,5\}$; moving-average kernels $\mathcal{K} = \{3,5,7\}$; maximum epoch $E$; early-stopping patience $p$

**Output:** Forecasts and evaluation scores for each forecasting horizon

Arrange the observations in chronological order and impute missing entries by time-based interpolation. When isolated missing values remain, apply forward and backward filling. Non-numeric fields that cannot be processed by the training pipeline are excluded from the station-level files;
**for each** *forecasting horizon* $h \in \mathcal{H}$ **do**
  Convert the time series into input–target pairs $(X_{t-L+1:t}, Y_{t+h})$ and split them chronologically into training, validation, and testing subsets using the 70:15:15 ratio;
  Initialize the parameters of the MK-DAF-LSTM model for horizon $h$. The best validation loss is recorded during training, and the patience counter is set to zero;
  **for** $e \leftarrow 1$ **to** $E$ **do**
    **for each** *training mini-batch* $(X, Y)$ **do**
      The raw input sequence is first passed through the base LSTM branch to obtain a direct recurrent forecast $\hat{Y}_{base}$;
      **for each** *kernel* $k_m \in \mathcal{K}$ **do**
        Apply moving-average decomposition to obtain the trend and residual components $(T^{(m)}, R^{(m)})$;
      Combine the trend components into $T_{mix}$ and the residual components into $R_{mix}$;
      Project $T_{mix}$ into a trend feature $z_T$. Encode $R_{mix}$ using the residual LSTM, followed by temporal attention, to obtain a residual feature $z_R$;

Estimate a fusion gate from the trend and residual features, then form the correction feature $z_c$;
Produce the correction term $\Delta\hat{Y}$ from $z_c$ and combine it with the base forecast using a conservative correction weight $\alpha$:

$$\hat{Y} = \hat{Y}_{\text{base}} + \alpha\Delta\hat{Y}.$$

Update the model parameters by minimizing the mean squared error between $Y$ and $\hat{Y}$ using the Adam optimizer;
Evaluate the model on the validation subset;
**if** *the validation loss improves* **then**
Store the current model as the best model for horizon $h$ and reset the patience counter;
**else**
Increase the patience counter by one;
**if** *the patience counter reaches* $p$ **then**
Stop the training process for the current horizon;

Restore the best model obtained for horizon $h$ and generate forecasts on the testing subset;
Compute MAE, RMSE, MAPE, and $R^2$ for the testing forecasts;
**return** *Forecasts and evaluation scores for all forecasting horizons*;

### 2.6. Datasets and preprocessing

The experiments used two publicly available Beijing air-quality datasets. The Beijing PM2.5 post-retirement studies of air quality (PRSA) dataset was used as the main benchmark for model comparison, repeat-seed evaluation, and ablation analysis. The Beijing multi-site air quality dataset was used as a supplementary station-wise robustness evaluation to examine whether the model behavior remains consistent across different monitoring locations. In this setting, four stations were selected: Aotizhongxin, Changping, Dongsi, and Huairou.

It should be noted that the multi-site evaluation in this study is station-wise. Each station was prepared and modeled as an independent time series. Therefore, the results are interpreted as evidence of cross-station robustness rather than as evidence of explicit spatiotemporal dependency modeling. This distinction is important because graph-based and multi-site air-quality forecasting models usually introduce explicit mechanisms to represent spatial relationships among monitoring stations. [18] used spatiotemporal fusion for multi-site PM2.5 prediction, Liu and Li [19] used graph convolution with LSTM, Wu *et al.* [21] used spatiotemporal graph attention, and Abbasi *et al.* [22] used clustered monitoring networks and pollutant-dependency information.

All records were first arranged in chronological order. Missing observations were imputed using time-based interpolation. Forward filling and backward filling were then applied only when interpolation did not fully resolve short gaps in the sequence. For station-level training files, variables that could not be represented numerically in the learning pipeline were excluded. The prediction target was PM2.5 concentration, represented as `PM25` in the prepared datasets. Table 1 summarizes the datasets used in the experiment, including their role, period, number of stations, and specific use in this study.

Table 1. Summary of datasets used in the experiment

| Dataset | Role | Period | Stations | Use in this study |
|---|---|---|---|---|
| Beijing PM2.5 single-site PRSA | Main dataset | 2010–2014 | 1 | Main comparison, repeat-seed analysis, and ablation study |
| Beijing Multi-Site Air Quality | Supplementary dataset | 2013–2017 | 12 available; 4 used | Station-wise cross-station robustness evaluation |

### 2.7. Experimental setup

The proposed model was compared with several baselines that represent the main families of sequence models used in short-term air-quality forecasting. LSTM, GRU, and BiLSTM were included to represent recurrent models. CNN-LSTM was used to represent convolutional-recurrent learning, following the common use of CNN-LSTM structures in PM2.5 forecasting such as the attention-based model of Li *et al.* [5]. Attention-LSTM was included because recent PM2.5 forecasting studies have shown that attention can help recurrent models emphasize more informative temporal patterns. This choice is consistent with the hybrid forecasting designs reported by Wei *et al.* [6] and Kumar [7]. DAF-LSTM was included to measure the benefit of the proposed multi-kernel extension over the single decomposition-aware variant. Each model received a 24-hour historical window and produced forecasts for 1-, 3-, and 5-hour horizons.

The same training setting was used across the compared models so that the reported differences were mainly attributable to the model structure rather than to different optimization choices. The main hyperparameters were fixed before the final comparison based on the validation setting and then kept consistent across horizons and baselines. The recurrent layers used 64 hidden units and two stacked layers. A dropout rate of 0.12 was applied to reduce overfitting, while the learning rate and batch size were set to 0.0008 and 64, respectively. Training was allowed to continue for up to 50 epochs. In practice, the best checkpoint was selected from the validation loss, and training was stopped when the validation performance did not improve for 10 consecutive epochs. All models were optimized with Adam using mean squared error as the training objective. Table 2 summarizes the configuration used in the experiments.

Table 2. Experimental configuration

| Component | Configuration |
|---|---|
| Proposed model | MK-DAF-LSTM |
| Baselines | LSTM, GRU, BiLSTM, CNN-LSTM, DAF-LSTM, attention-LSTM |
| Input sequence length | 24 hours |
| Forecasting horizons | 1, 3, and 5 hours |
| Moving-average kernels | 3, 5, and 7-time steps |
| Correction scale | Learnable $\alpha = \sigma(\beta)$ with $\beta$ initialized to $-3.0$ |
| Hidden dimension | 64 |
| Number of LSTM layers | 2 |
| Dropout | 0.12 |
| Learning rate | 0.0008 |
| Batch size | 64 |
| Epochs | 50 |
| Early stopping patience | 10 |
| Optimizer | Adam |
| Evaluation metrics | Mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and $R^2$ |
| Repeat-seed protocol | Three independent seeds: 42, 123, and 2026 |
| Multi-site setting | Each station is treated as an independent time series |

### 2.8. Implementation details and reproducibility protocol

The implementation used the MK-DAF-LSTM architecture described in section 2. The moving-average decomposition used replicate padding and the fixed kernel set {3,5,7}. The learnable kernel selector was implemented as a two-layer feed-forward network with rectified linear unit (ReLU) activation followed by a softmax over the three kernels. The temporal attention module used a linear-tanh-linear scoring network followed by softmax normalization over the time dimension. The fusion gate used a two-layer feed-forward network with sigmoid output, and the conservative correction scale was initialized through $\beta = -3.0$ in $\alpha = \sigma(\beta)$. Except for the correction scale, all other trainable layers used the default PyTorch initialization.

The repeat-seed evaluation used three independent random seeds: 42, 123, and 2026. Results from this evaluation are reported as mean $\pm$ standard deviation. The computational environment used Python 3.10.12, PyTorch 2.3.1+cu121, CUDA 12.1, and cuDNN 8902 on a CUDA-enabled Linux workstation equipped with two NVIDIA A100 80 GB PCIe GPUs. For the single-site MK-DAF-LSTM configuration, the input shape was $[64,24,15]$, corresponding to batch size 64, a 24-hour input window, and 15 input features. The trained model checkpoints contained 102,087 trainable parameters. Inference efficiency was measured using trained MK-DAF-LSTM checkpoints with 100 warm-up runs and 1,000 timed forward-pass runs. Across the three 3-hour repeat-seed checkpoints, the mean inference time was $1.5056 \pm 0.0141$ ms per batch, equivalent to $0.0235 \pm 0.0002$ ms per sample.

### 2.9. Evaluation metrics

Model accuracy was assessed from both absolute and relative error perspectives. MAE was used to describe the average magnitude of the forecasting error, whereas RMSE was included because large errors are more harmful in air-quality warning applications. MAPE was reported to provide a relative error view, and $R^2$ was used to indicate how much variation in the observed PM2.5 concentration could be explained by the model. The four metrics are defined as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right| \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2} \tag{14}$$

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{Y_i-\hat{Y}_i}{Y_i+\epsilon}\right| \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i-\bar{Y})^2} \quad (16)$$

where $Y_i$ is the observed PM2.5 concentration, $\hat{Y}_i$ is the predicted value, $\bar{Y}$ is the mean of the observed testing values, and $N$ is the number of testing samples. A small constant $\epsilon$ is added in the MAPE calculation to avoid division by zero.

In interpreting the results, MAE and RMSE were treated as the primary error measures, while $R^2$ was used to assess explained variance. MAPE was retained as a complementary indicator because percentage errors can be inflated when the observed PM2.5 concentration is close to zero.

## 3. RESULTS AND DISCUSSION

This section reports the performance of MK-DAF-LSTM from four viewpoints. The first analysis compares the proposed model with recurrent and hybrid baselines on the Beijing PM2.5 single-site dataset. The second analysis examines whether the performance remains stable across repeated training runs. The third analysis uses ablation experiments to evaluate the contribution of the multi-kernel decomposition-aware fusion mechanism. The fourth analysis uses selected stations from the Beijing multi-site air quality dataset to examine station-wise robustness. Since the multi-site data are modeled independently for each station, the results should be interpreted as cross-station robustness rather than full spatiotemporal dependency modeling.

### 3.1. Main forecasting performance on the single-site dataset

Table 3 presents the forecasting results on the Beijing PM2.5 single-site dataset for 1-, 3-, and 5-hour horizons. The compared baselines were selected to reflect model families that are commonly used in recent PM2.5 forecasting studies. CNN-LSTM with attention, hybrid CNN-LSTM-attention models, and recurrent attention-based predictors have been widely used in this area [5]–[7]. A decomposition-hybrid reference was also included because decomposition has been used before CNN-BiLSTM-attention prediction [13]. For ease of reading, the best value for each metric at a given forecasting horizon is shown in bold, while the second-best value is underlined. Lower MAE, RMSE, and MAPE indicate smaller forecasting errors, whereas higher $R^2$ indicates stronger agreement with the observed PM2.5 series.

Table 3. Single-site forecasting results on the Beijing PM2.5 dataset

| Model | Horizon | MAE | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| MK-DAF-LSTM | 1 hour | 11.6521 | 20.3187 | 23.0069 | 0.9531 |
| LSTM | 1 hour | 11.4605 | 20.3332 | 21.9578 | 0.9530 |
| Attention-LSTM | 1 hour | 11.5050 | 20.3351 | 20.8741 | 0.9530 |
| DAF-LSTM | 1 hour | 11.6484 | 20.4068 | 22.9713 | 0.9527 |
| CNN-LSTM | 1 hour | 11.7149 | 20.6053 | 22.5326 | 0.9518 |
| BiLSTM | 1 hour | 11.6052 | 20.7505 | 21.2046 | 0.9511 |
| GRU | 1 hour | 11.9179 | 21.1137 | 22.9937 | 0.9494 |
| MK-DAF-LSTM | 3 hours | 23.1872 | 37.4176 | 46.4904 | 0.8409 |
| Attention-LSTM | 3 hours | 23.1062 | 37.4417 | 48.0176 | 0.8407 |
| DAF-LSTM | 3 hours | 23.6336 | 37.8495 | 51.6265 | 0.8373 |
| GRU | 3 hours | 23.3406 | 37.8598 | 52.4779 | 0.8372 |
| CNN-LSTM | 3 hours | 23.6124 | 37.9311 | 53.2593 | 0.8365 |
| LSTM | 3 hours | 24.1153 | 39.1048 | 48.9369 | 0.8263 |
| BiLSTM | 3 hours | 24.2028 | 39.4855 | 49.0430 | 0.8229 |
| Attention-LSTM | 5 hours | 30.7139 | 47.5878 | 75.5083 | 0.7427 |
| MK-DAF-LSTM | 5 hours | 31.0029 | 47.9327 | 79.0459 | 0.7390 |
| DAF-LSTM | 5 hours | 30.7223 | 48.1540 | 69.8973 | 0.7366 |
| CNN-LSTM | 5 hours | 31.6376 | 48.7468 | 79.4537 | 0.7300 |
| BiLSTM | 5 hours | 32.8169 | 49.4307 | 85.7954 | 0.7224 |
| GRU | 5 hours | 32.2956 | 49.5740 | 73.7237 | 0.7208 |
| LSTM | 5 hours | 32.6949 | 49.6007 | 82.4573 | 0.7205 |

For the 1-hour horizon, the comparison is very close. LSTM gives the lowest MAE, and attention-LSTM gives the lowest MAPE. MK-DAF-LSTM, however, obtains the lowest RMSE and the highest $R^2$. This pattern suggests that the proposed model is not necessarily the best in average absolute error for the very

short horizon, but it is slightly better at reducing larger deviations, which are more strongly reflected by RMSE.

The 3-hour horizon gives the clearest advantage for MK-DAF-LSTM. Although attention-LSTM has a slightly lower MAE, MK-DAF-LSTM obtains the best RMSE, MAPE, and $R^2$. This result is important because the 3-hour setting is no longer a purely immediate forecast, but it still depends strongly on recent temporal information. The result supports the use of decomposition-aware temporal representation, especially when the model needs to represent both local fluctuation and smoother temporal behavior. This interpretation is consistent with Zaini *et al.* [10], who showed that decomposition can support PM2.5 forecasting. It is also consistent with Dong *et al.* [14] who emphasized multi-scale feature extraction, and Xu *et al.* [15], who used a multi-kernel decomposition strategy.

For the 5-hour horizon, attention-LSTM gives the best MAE, RMSE, and $R^2$ in the single-run comparison, while DAF-LSTM gives the lowest MAPE. MK-DAF-LSTM is therefore not the single-run winner for this horizon. Nevertheless, it remains in the top-performing group, particularly for RMSE and $R^2$. This finding should be read together with the repeat-seed analysis in Table 4, where the proposed model shows a more stable average result for the longer horizon.

### 3.2. Repeat-seed stability

Single-run results may be affected by initialization and training dynamics. To reduce this risk, repeat-seed evaluation was carried out using three independent random seeds: 42, 123, and 2026. The evaluation was applied to the horizons where the proposed model and the strongest baselines were close. The summary in Table 4 reports the mean and standard deviation across repeated runs.

Table 4. Repeat-seed stability summary on the Beijing PM2.5 single-site dataset

| Model | Horizon | MAE | RMSE | MAPE | $R^2$ | Parameters |
|---|---|---|---|---|---|---|
| MK-DAF-LSTM | 1 hour | 11.6939±0.1588 | 20.3786±0.0663 | 23.2863±1.1309 | 0.9528±0.0003 | 102,087 |
| MK-DAF-LSTM | 3 hours | 23.2834±0.1810 | 37.5933±0.1971 | 48.7795±2.0075 | 0.8394±0.0017 | 102,087 |
| Attention-LSTM | 3 hours | 23.4098±0.3106 | 37.8476±0.5472 | 50.2485±3.0064 | 0.8372±0.0047 | 62,466 |
| MK-DAF-LSTM | 5 hours | 31.3448±0.2974 | 48.2680±0.3141 | 76.6855±2.0676 | 0.7353±0.0034 | 102,087 |
| DAF-LSTM | 5 hours | 31.2316±0.5301 | 48.6833±0.4680 | 71.3144±3.8247 | 0.7307±0.0052 | 100,868 |
| Attention-LSTM | 5 hours | 31.8120±1.4908 | 48.8472±1.4886 | 75.3538±8.2170 | 0.7288±0.0166 | 62,466 |

The repeat-seed results further support the performance of MK-DAF-LSTM at the medium and longer horizons. For the 3-hour horizon, MK-DAF-LSTM achieves lower average MAE, RMSE, and MAPE than attention-LSTM, while also producing a higher average $R^2$ and a smaller standard deviation in RMSE. For the 5-hour horizon, DAF-LSTM still gives slightly lower average MAE and MAPE, but MK-DAF-LSTM obtains the best average RMSE and $R^2$. This pattern is consistent with the main comparison and shows that the proposed model is stronger in reducing larger errors and preserving explained variance than in improving every error measure equally.

The stability analysis also clarifies the interpretation of the 5-hour single-run result. Although attention-LSTM is the best model in the single-run 5-hour comparison, its repeat-seed average becomes less favorable than MK-DAF-LSTM in RMSE and $R^2$. Thus, the safer conclusion is that MK-DAF-LSTM provides more stable longer-horizon behavior in the main error-sensitive metrics, rather than uniformly outperforming all baselines.

### 3.3. Paired uncertainty analysis

Because several performance differences are numerically small, an additional paired uncertainty analysis was conducted using the stored test-set predictions. The analysis used a moving-block bootstrap with 24-hour blocks on the squared-error difference between MK-DAF-LSTM and the strongest competing baselines. The difference is defined as the baseline squared error minus the MK-DAF-LSTM squared error, so a positive value indicates a lower squared error for MK-DAF-LSTM. Table 5 reports the mean difference, 95% bootstrap confidence interval, and two-sided bootstrap probability.

The paired analysis confirms that several differences are within the expected uncertainty range, especially for the 1-hour horizon and the close comparison against attention-LSTM at the 3-hour horizon. Therefore, the statistical evidence is used to calibrate the interpretation rather than to claim universal superiority. The revised claim is that MK-DAF-LSTM is competitive in single-run comparison and more stable in repeat-seed RMSE and $R^2$, particularly at the 3- and 5-hour horizons.

Table 5. Paired uncertainty analysis based on test-set squared-error differences

| Horizon | Baseline compared with MK-DAF-LSTM | Mean difference | 95% block-bootstrap CI | $p$ |
|---|---|---|---|---|
| 1 hour | LSTM | 0.5893 | [−7.1643, 8.6314] | 0.904 |
| 3 hours | Attention-LSTM | 1.8000 | [−54.0979, 55.6852] | 0.974 |
| 3 hours | DAF-LSTM | 32.5059 | [−1.2222, 68.3684] | 0.061 |
| 5 hours | Attention-LSTM | −32.9531 | [−138.3608, 62.9539] | 0.539 |
| 5 hours | DAF-LSTM | 21.2587 | [−62.0520, 119.1134] | 0.678 |

### 3.4. Ablation study

The ablation study was used to check whether the performance gain is associated with the complete multi-kernel decomposition-aware design. Table 6 compares the full MK-DAF-LSTM with the single-kernel DAF-LSTM, the equal-weight multi-kernel variant, and the LSTM baseline. The full MK-DAF-LSTM gives the best RMSE and $R^2$ for both 3- and 5-hour horizons. At the 3-hour horizon, it also gives the best MAE among the ablation variants, although the LSTM baseline has a lower MAPE. At the 5-hour horizon, DAF-LSTM and the equal-weight variant have lower MAE and MAPE, but both are worse than the full MK-DAF-LSTM in RMSE and $R^2$. This confirms that the main benefit of the full model is not a uniform reduction in all metrics, but a more consistent reduction of larger errors and better explained variance.

The ablation result supports the role of the learnable multi-kernel correction branch. The full model performs better than the equal-weight variant in RMSE and $R^2$, indicating that learning how to combine temporal views is more effective than assigning the same weight to every kernel. The result also shows that the single-kernel decomposition-aware model is useful, but not as strong as the full multi-kernel version in the main metrics. This interpretation is consistent with prior decomposition-based and multi-scale PM2.5 forecasting studies [10], [14]–[16]. The ablation was designed to compare kernel-use strategies rather than to perform an exhaustive search over all possible kernel combinations. Therefore, the selected kernel set {3, 5, 7} should be interpreted as a compact and practically motivated configuration for the 24-hour input window.

Table 6. Ablation study of MK-DAF-LSTM on the Beijing PM2.5 single-site dataset

| Variant | Horizon | MAE | RMSE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| MK-DAF-LSTM full | 3 hours | 23.1708 | 37.5557 | 49.6084 | 0.8398 |
| DAF-LSTM single-kernel | 3 hours | 23.6336 | 37.8495 | 51.6265 | 0.8373 |
| MK-DAF-LSTM equal-weight | 3 hours | 24.2050 | 38.2461 | 56.1732 | 0.8338 |
| LSTM baseline | 3 hours | 24.1153 | 39.1048 | 48.9369 | 0.8263 |
| MK-DAF-LSTM full | 5 hours | 31.0029 | 47.9327 | 79.0459 | 0.7390 |
| DAF-LSTM single-kernel | 5 hours | 30.7223 | 48.1540 | 69.8973 | 0.7366 |
| MK-DAF-LSTM equal-weight | 5 hours | 30.8231 | 48.3081 | 70.1530 | 0.7349 |
| LSTM baseline | 5 hours | 32.6949 | 49.6007 | 82.4573 | 0.7205 |

### 3.5. Computational profile

The additional correction branch increases the parameter count of MK-DAF-LSTM relative to standard LSTM. Table 7 summarizes the trainable parameter counts and training times for MK-DAF-LSTM and representative strong baselines. The training time should be interpreted together with early stopping because different models may stop at different epochs.

These results show that MK-DAF-LSTM is more complex than standard LSTM because it adds multi-kernel decomposition, residual recurrent encoding, temporal attention, and a correction head. However, the trained-checkpoint inference benchmark indicates that the forward-pass cost remains low. Across the three 3-hour repeat-seed checkpoints, MK-DAF-LSTM required $1.5056 \pm 0.0141$ ms per batch and $0.0235 \pm 0.0002$ ms per sample for the input shape [64, 24, 15]. The model should therefore be interpreted as an accuracy–stability trade-off rather than as the most computationally efficient baseline.

Table 7. Computational profile of MK-DAF-LSTM and representative baselines

| Model | Horizon | Parameters | Train time (s) |
|---|---|---|---|
| MK-DAF-LSTM | 1 hour | 102,087 | 210.40 |
| LSTM | 1 hour | 58,241 | 48.49 |
| Attention-LSTM | 1 hour | 62,466 | 104.14 |
| MK-DAF-LSTM | 3 hours | 102,087 | 155.33 |
| Attention-LSTM | 3 hours | 62,466 | 65.05 |
| DAF-LSTM | 3 hours | 100,868 | 106.86 |
| MK-DAF-LSTM | 5 hours | 102,087 | 79.19 |
| Attention-LSTM | 5 hours | 62,466 | 74.01 |
| DAF-LSTM | 5 hours | 100,868 | 107.57 |

### 3.6. Station-wise robustness on the multi-site dataset

The multi-site experiment was used as a supplementary evaluation. Four stations from the Beijing multi-site air quality dataset were selected: Aotizhongxin, Changping, Dongsi, and Huairou. Each station was trained and evaluated independently. Table 8 reports the average station-wise performance across the four stations.

Table 8 summarizes the average results obtained from the four selected stations. The 1-hour forecast gives the strongest station-wise result, with an average $R^2$ of 0.9509. The value decreases to 0.8061 at the 3-hour horizon and to 0.6930 at the 5-hour horizon. This decline is reasonable because longer horizons depend less directly on the most recent observations and are therefore harder to predict.

The station-wise results indicate that MK-DAF-LSTM can be applied consistently across the selected monitoring locations, especially for short-term forecasting. This finding should still be interpreted carefully. The experiment does not test whether the model learns spatial dependency among stations. Wang *et al.* [18] addressed this issue through spatiotemporal fusion, while Liu and Li [19] used graph convolution to represent relationships between monitoring stations. Wu *et al.* [21] further used spatiotemporal graph attention with multi-source data, and Abbasi *et al.* [22] explored clustered monitoring networks and pollutant dependency. In contrast, the present study treats each station as an independent time series. Therefore, the claim is limited to station-wise robustness.

Table 8. Average forecasting performance across the selected Beijing Multi-Site stations

| Horizon | MAE | RMSE | MAPE | Average $R^2$ | Avg train time (s) |
|---|---|---|---|---|---|
| 1 hour | 10.1519 | 19.2974 | 29.4501 | 0.9509 | 117.38 |
| 3 hours | 21.4651 | 38.4897 | 67.7516 | 0.8061 | 120.58 |
| 5 hours | 28.5378 | 48.6628 | 97.2445 | 0.6930 | 91.71 |

### 3.7. Interpretation and limitations

The overall results indicate that MK-DAF-LSTM is most useful when the forecasting task benefits from a richer temporal representation. The horizon-wise pattern provides a useful interpretation of the correction mechanism. At the 1-hour horizon, the most recent observations are still highly informative, so a standard LSTM remains very competitive in MAE and attention-LSTM remains strong in MAPE. In this setting, MK-DAF-LSTM mainly improves RMSE and $R^2$, suggesting that the correction branch helps reduce larger deviations rather than uniformly reducing all error measures. At the 3-hour horizon, the task is far enough from the observed window to benefit from trend-residual temporal views, but not so far that the recent temporal structure becomes too weak. This explains why the 3-hour horizon provides the clearest support for the proposed multi-kernel correction branch. At the 5-hour horizon, uncertainty increases and local temporal information becomes less directly predictive. Consequently, attention-LSTM remains strong in the single-run comparison, while repeat-seed evaluation shows that MK-DAF-LSTM provides more stable RMSE and $R^2$ behavior.

Several limitations should be stated explicitly. First, the proposed model does not outperform all baselines on every metric. Its advantage is mainly observed in RMSE, $R^2$, and repeat-seed stability. Second, the multi-site experiment is station-wise and does not include explicit spatial dependency modeling. Third, the model is more complex than a standard LSTM because it adds a multi-kernel decomposition-aware correction branch. Fourth, MAPE should be interpreted carefully because percentage errors can be inflated when the observed PM2.5 concentration is small. Finally, this study is limited to forecasting accuracy, stability, and ablation-based validation of the proposed mechanism. Future work should examine the model more closely under practical air-quality conditions. One direction is to relate the forecasts to meteorological and pollutant factors, since such variables can provide useful explanatory context for PM2.5 variation [25]. Another direction is to revisit the temporal setting itself. Air-pollution changes are not always evenly paced, and continuous-time or multi-timescale forecasting may offer a better way to describe such irregular behavior [26]. Future studies may also extend the present station-wise setting by adding explicit spatial information, multimodal observations, or satellite-derived signals, following recent developments in spatiotemporal and multimodal air-pollution forecasting [27], [28]

Overall, the results show that MK-DAF-LSTM offers competitive and stable PM2.5 forecasting performance by combining a base recurrent forecast with a multi-kernel decomposition-aware correction branch. The clearest support appears at the 3-hour horizon, where the model performs strongly in the main comparison. The longer-horizon results further indicate that the proposed model is useful when stability in RMSE and explained variance is more important than improvement in every error metric.

## 4. CONCLUSION

This study proposed MK-DAF-LSTM, a multi-kernel decomposition-aware recurrent fusion model for short-term PM2.5 forecasting. The model combines a base recurrent forecast with a learnable correction branch derived from moving-average trend-residual representations. The kernel set {3,5,7}, learnable kernel selector, temporal attention module, fusion gate, and conservative correction scale allow the model to use complementary temporal views while preserving the stability of the base LSTM forecast. Experiments on the Beijing PM2.5 single-site dataset show that MK-DAF-LSTM is competitive against recurrent and hybrid baselines across 1-, 3-, and 5-hour forecasting horizons. The strongest evidence is observed at the 3-hour horizon, where MK-DAF-LSTM achieves the best RMSE, MAPE, and R2 in the main comparison. Repeat-seed evaluation using three random seeds further shows that the proposed model provides lower average RMSE and higher average R2 than the strongest repeated baselines at the 3- and 5-hour horizons. The paired uncertainty analysis also supports a conservative interpretation: several numerical differences are small, so the contribution should be read as improved competitiveness and stability rather than universal superiority. The ablation study confirms that the full multi-kernel design is more effective than the single-kernel DAF-LSTM and the equal-weight multi-kernel variant in the main error-sensitive metrics, especially RMSE and R2. The supplementary evaluation on selected Beijing Multi-Site stations further shows that the model maintains strong station-wise performance at the 1-hour horizon, although performance decreases as the forecasting horizon becomes longer. These findings indicate that MK-DAF-LSTM can support practical short-term air-quality monitoring and early warning by providing a compact recurrent model that remains stable across repeated training runs. The proposed model should not be interpreted as a universal winner across all metrics or as a full spatiotemporal forecasting model. Future work may extend the present design by incorporating graph-based spatial dependency modeling, uncertainty quantification for high-pollution episodes, interpretable environmental factor analysis, and operational deployment evaluation in real air-quality monitoring systems.

## ACKNOWLEDGMENTS

The authors thank Universitas Negeri Semarang for institutional support during the preparation of this study. The authors also thank the providers of the public Beijing PM2.5 and Beijing multi-site air quality datasets used in the experiments.

## FUNDING INFORMATION

This research was funded by the Institute for Research and Community Service (LPPM), Universitas Negeri Semarang, through the 2024 Budget Implementation List (DPA LPPM UNNES), under DPA number 023.17.2.690645/2024.10 and contract number 82.26.2/UN37/PPK.10/2024.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wahyu Hardyanto | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | | ✓ | | |
| Siti Wahyuni | | ✓ | | ✓ | | ✓ | | | | ✓ | | ✓ | | |
| Aji Purwinarko | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | |
| Isa Akhlis | | ✓ | | ✓ | | ✓ | | | | ✓ | | ✓ | | |
| Mahargjo Hapsoro Adi | | | ✓ | | | ✓ | | ✓ | ✓ | ✓ | ✓ | | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

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

## DATA AVAILABILITY

The raw datasets used in this study are publicly available Beijing air-quality datasets. The Beijing PM2.5 single-site PRSA dataset was used for the main model comparison, repeat-seed evaluation, and ablation analysis. The Beijing multi-site air quality dataset was used for supplementary station-wise robustness evaluation. The processed data, experiment logs, and model outputs can be made available from the corresponding author upon reasonable request. If a public repository is prepared before final publication, the repository link will be added to the final manuscript.

## REFERENCES


[1] X. Feng *et al.*, "A hybrid model for enhanced forecasting of PM2.5 spatiotemporal concentrations with high resolution and accuracy," *Environmental Pollution*, vol. 355, p. 124263, 2024, doi: 10.1016/j.envpol.2024.124263.

[2] S. Gul, G. M. Khan, and S. Yousaf, "Multi-step short-term PM2.5 forecasting for enactment of proactive environmental regulation strategies," *Environmental Monitoring and Assessment*, vol. 194, no. 5, 2022, doi: 10.1007/s10661-022-10029-4.

[3] M. S. Qamar, M. F. Munir, and A. Waseem, "AI for Cleaner Air: Predictive Modeling of PM2.5 Using Deep Learning and Traditional Time-Series Approaches," *Computer Modeling in Engineering & Sciences*, vol. 144, no. 3, pp. 3557–3584, 2025, doi: 10.32604/cmes.2025.067447.

[4] C.-Y. Lee, J.-Y. Lee, S.-H. Han, J.-G. Kang, J.-B. Lee, and D.-R. Choi, "Performance Evaluation of PM2.5 Forecasting Using SARIMAX and LSTM in the Korean Peninsula," *Atmosphere*, vol. 16, no. 5, p. 524, 2025, doi: 10.3390/atmos16050524.

[5] S. Li, G. Xie, J. Ren, L. Guo, Y. Yang, and X. Xu, "Urban PM2.5 Concentration Prediction via Attention-Based CNN–LSTM," *Applied Sciences*, vol. 10, no. 6, p. 1953, 2020, doi: 10.3390/app10061953.

[6] Q. Wei, H. Zhang, J. Yang, B. Niu, and Z. Xu, "PM2.5 concentration prediction using a whale optimization algorithm based hybrid deep learning model in Beijing, China," *Environmental Pollution*, vol. 371, p. 125953, 2025, doi: 10.1016/j.envpol.2025.125953.

[7] S. Kumar, V. Kour, P. Kumar, A. Deshmane, S. Mishra, and R. Misra, "Gru-lstm with attention-based forecasting for enhanced air quality," *Natural Hazards*, vol. 121, no. 13, pp. 15925–15947, 2025, doi: 10.1007/s11069-025-07408-8.

[8] A. Aggarwal and D. Toshniwal, "A hybrid deep learning framework for urban air quality forecasting," *Journal of Cleaner Production*, vol. 329, p. 129660, 2021, doi: 10.1016/j.jclepro.2021.129660.

[9] Z. Zhang, Y. Zeng, and K. Yan, "A hybrid deep learning technology for PM2.5 air quality forecasting," *Environmental Science and Pollution Research*, vol. 28, no. 29, pp. 39409–39422, 2021, doi: 10.1007/s11356-021-12657-8.

[10] N. Zaini, L. W. Ean, A. N. Ahmed, M. A. Malek, and M. F. Chow, "PM2.5 forecasting for an urban area based on deep learning and decomposition method," *Scientific Reports*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-21769-1.

[11] Q. Zheng *et al.*, "Application of wavelet-packet transform driven deep learning method in PM2.5 concentration prediction: A case study of Qingdao, China," *Sustainable Cities and Society*, vol. 92, p. 104486, 2023, doi: 10.1016/j.scs.2023.104486.

[12] S. Li, B. Tang, and X. Deng, "A Hybrid Method Combining Variational Mode Decomposition and Deep Neural Networks for Predicting PM2.5 Concentration in China," *IEEE Access*, vol. 13, pp. 51956–51968, 2025, doi: 10.1109/access.2025.3552229.

[13] T. Sreenivasulu and G. M. Rayalu, "Enhanced PM2.5 prediction in Delhi using a novel optimized STL-CNN-BILSTM-AM hybrid model," *Asian Journal of Atmospheric Environment*, vol. 18, no. 1, 2024, doi: 10.1007/s44273-024-00048-7.

[14] L. Dong, P. Hua, D. Gui, and J. Zhang, "Extraction of multi-scale features enhances the deep learning-based daily PM2.5 forecasting in cities," *Chemosphere*, vol. 308, p. 136252, 2022, doi: 10.1016/j.chemosphere.2022.136252.

[15] L. Xu, J.-H. Zhang, and C.-W. Yin, "PM2.5 prediction using hybrid relevance vector machine with ISSA-optimized multi-kernel and variational mode decomposition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 17, no. 1, pp. 307–323, 2025, doi: 10.1007/s12652-025-04985-1.

[16] X. Guan, X. Mo, H. Li, Q. Dong, and Z. Zhang, "A multi-step forecasting framework for PM $_{2.5}$ based on TVFEMD-OVMD decomposition and ALO-optimized hybrid deep learning network," *Engineering Research Express*, vol. 8, no. 7, p. 75243, 2026, doi: 10.1088/2631-8695/ae5b3f.

[17] U. Pak *et al.*, "Novel particulate matter (PM2.5) forecasting method based on deep learning with suitable spatiotemporal correlation analysis," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 264, p. 106336, 2024, doi: 10.1016/j.jastp.2024.106336.

[18] H. Wang, L. Zhang, R. Wu, and Y. Cen, "Spatio-temporal fusion of meteorological factors for multi-site PM2.5 prediction: A deep learning and time-variant graph approach," *Environmental Research*, vol. 239, p. 117286, 2023, doi: 10.1016/j.envres.2023.117286.

[19] X. Liu and W. Li, "MGC-LSTM: a deep learning model based on graph convolution of multiple graphs for PM2.5 prediction," *International Journal of Environmental Science and Technology*, vol. 20, no. 9, pp. 10297–10312, 2022, doi: 10.1007/s13762-022-04553-6.

[20] Y. Chen *et al.*, "Regional PM2.5 prediction with hybrid directed graph neural networks and Spatio-temporal fusion of meteorological factors," *Environmental Pollution*, vol. 366, p. 125404, 2025, doi: 10.1016/j.envpol.2024.125404.

[21] J. Wu, L. Xu, S. Fan, K. Xia, and L. Wang, "A Spatiotemporal graph attention network for PM2.5 forecasting using multi-source data," *Air Quality, Atmosphere & Health*, vol. 18, no. 10, pp. 3037–3052, 2025, doi: 10.1007/s11869-025-01818-0.

[22] M. T. Abbasi, A. A. Alesheikh, and F. Rezaie, "A Lightweight Spatiotemporal Graph Framework Leveraging Clustered Monitoring Networks and Copula-Based Pollutant Dependency for PM2.5 Forecasting," *Land*, vol. 14, no. 8, p. 1589, 2025, doi: 10.3390/land14081589.

[23] Y. Hu, Q. Li, X. Shi, J. Yan, and Y. Chen, "Domain knowledge-enhanced multi-spatial multi-temporal PM2.5 forecasting with integrated monitoring and reanalysis data," *Environment International*, vol. 192, p. 108997, 2024, doi: 10.1016/j.envint.2024.108997.

[24] L. Zhang, J. Na, J. Zhu, Z. Shi, C. Zou, and L. Yang, "Spatiotemporal causal convolutional network for forecasting hourly PM2.5 concentrations in Beijing, China," *Computers & Geosciences*, vol. 155, p. 104869, 2021, doi: 10.1016/j.cageo.2021.104869.

[25] D. Li, M. Liu, H. Han, and J. Wang, "Nonlinear Impacts of Air Pollutants and Meteorological Factors on PM2.5: An Interpretable GT-iFormer Model with SHAP Analysis," *Atmosphere*, vol. 17, no. 3, p. 266, 2026, doi: 10.3390/atmos17030266.

[26] M. A. Havaei, V. Shahhosseini, and R. Maknoon, "Continuous-time air pollutant forecasting using multi-timescale attention neural ordinary differential equations (MA-NODE)," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-27903-z.

[27] F. Zhang, J. Hu, and M. Zeng, "A Spatiotemporal Multimodal Framework for Air Pollution Prediction Based on Bayesian

Optimization—Evidence from Sichuan, China," *Atmosphere*, vol. 16, no. 8, p. 958, 2025, doi: 10.3390/atmos16080958.
[28] T. Srimanee, B. Kiatphibun, C. Thanomngern, N. Mhuaksa, C. Banju-ngam, and T. Siriborvornratanakul, "Multimodal PM2.5 Forecasting Using Satellite Imagery and Sensor Data with Semi-supervised Deep Learning," *Earth Systems and Environment*, vol. 10, no. 8, pp. 10147–10164, 2026, doi: 10.1007/s41748-026-01109-3.

## BIOGRAPHIES OF AUTHORS



**Wahyu Hardyanto** is a Professor of Computer Modeling at Universitas Negeri Semarang, Indonesia. His research interests include computational physics, computer modelling, numerical simulation, forward and inverse modelling, machine learning, neural networks, and information technology for education and management. He has also contributed to the development of physics learning software and academic information systems. He can be contacted through his institutional email at Universitas Negeri Semarang. He can be contacted at email: hardy@mail.unnes.ac.id.



**Siti Wahyuni** is a researcher and faculty member in the Department of Physics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia. Her research interests include physics, physics education, astrophysics, mathematical physics, and technology-supported learning. Her recent publications cover physics learning, digital teaching materials, computational notebooks, scientific creativity, and educational innovation. She can be contacted through her institutional email at Universitas Negeri Semarang. She can be contacted at email: wahyuni.smg@mail.unnes.ac.id.



**Aji Purwinarko** is a lecturer in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia. His academic and research activities are mainly related to computer science, machine learning, data-driven modelling, information systems, and applied intelligent computing. His recent works include applications of machine learning and deep learning for sentiment analysis, stock prediction, medical data analysis, and learning media. He can be contacted at email: aji.purwinarko@mail.unnes.ac.id.



**Isa Akhlis** is a lecturer in the Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia. His academic interests include computer programming, image processing, physics education, digital learning media, and educational technology. His works include the development of digital practicum tools, physics learning media, and technology-based learning resources. He can be contacted at email: isa.akhlis@mail.unnes.ac.id.



**Mahargjo Hapsoro Adi** is affiliated with Universitas Negeri Semarang, Indonesia. His academic works are related to library services, library information systems, information retrieval services, Android-based library applications, library service quality, and academic library development. His publications include studies on library notification systems, Android-based information retrieval services, and library service quality in higher education contexts. He can be contacted through his institutional email at Universitas Negeri Semarang. He can be contacted at email: hapsoroadi@mail.unnes.ac.id.