

CNN-LSTM with Scaled Dot-Product Attention for PM2.5 Forecasting in Jakarta: A Univariate Baseline

Brilliant Hendrylie^{1*}, Puguh Hiskiawan^{2*}

* Data Science Program, Faculty of Technology and Design, Bunda Mulia University
s36230001@student.ubm.ac.id¹, phiskiawan@bundamulia.ac.id²

Article Info

Article history:

Received 2026-08-18

Revised 2026-09-11

Accepted 2026-10-05

Keyword:

Air Quality,
Attention Mechanism,
CNN-LSTM,
Deep Learning,
Ensemble Forecasting,
Univariate Baseline.

ABSTRACT

This study evaluates a CNN-LSTM architecture with scaled dot-product attention for hour-ahead PM2.5 forecasting in Jabodetabek. Due to data constraints, the final model operates on a single univariate input channel - PM2.5's own first-order lag ($t-1$) - rather than the intended multivariate design; a supplementary experiment that added cyclical calendar attributes (hour, day-of-week, month) to the input did not improve accuracy and was therefore excluded from the final architecture. A three-member ensemble achieves $R^2 = 0.788$ and $MAE = 9.654 \mu\text{g}/\text{m}^3$ on the held-out test set, outperforming a naive persistence baseline on R^2 (0.678) and RMSE, though the two are comparable on MAE, indicating the model's advantage lies chiefly in curbing large errors rather than typical-case error. Performance is strong in the dominant high-concentration regime but degrades sharply in the sparsely represented low-to-moderate range, and residual diagnostics indicate the model has not fully captured the series' temporal dependence. These results position the architecture as a working univariate baseline rather than an operationally validated forecaster, and motivate future work incorporating co-pollutant and meteorological data with per-city sequence construction to realize its full multivariate design.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

Air pollution, specifically the escalation of Particulate Matter 2.5 (PM2.5), remains a critical environmental and public health crisis in the Jakarta metropolitan area [1]. Due to its microscopic aerodynamic diameter, PM2.5 possesses high viability to penetrate the human alveolar barrier, exacerbating severe pulmonary and cardiovascular pathologies. Data-driven screening and monitoring tools, particularly hybrid deep learning frameworks, have increasingly been proposed as a complementary line of defense for this kind of public-health burden by bridging environmental exposure forecasting with healthcare impact mitigation [1],[2],[3]. In a dense urban agglomeration like Jakarta, PM2.5 concentrations are not static; they exhibit volatility driven by a complex, non-linear interplay of localized primary emissions, secondary aerosol formation, and meteorological dispersion factors such as wind velocity, temperature, and relative humidity [4]. Developing a precise predictive framework for PM2.5 dynamics is therefore an

operational imperative for proactive public health intervention and environmental policy [1].

Historically, air quality forecasting relied on classical statistical and shallow machine-learning approaches, which struggle to capture the non-linear, stochastic nature of urban air quality data. Deep learning - particularly Long Short-Term Memory (LSTM) networks - has since become a standard approach for sequential forecasting due to its capacity to model long-term temporal dependencies [5]-[8], a capacity that has likewise proven effective for other sequential and signal-based recognition tasks, such as neural-network-based classification of Indonesian traditional music from Mel-frequency cepstral features [9]. However, a purely temporal model is structurally limited in exploiting multivariate correlations - for instance, between PM2.5, PM10, and other co-pollutants - when such correlated variables are available [4], [10]. To address this, hybrid CNN-LSTM architectures have been proposed, in which one-dimensional convolutional layers act as feature extractors ahead of a recurrent temporal model [5], [10], [11], following

a broader trend in which convolutional networks have been adapted as general-purpose feature extractors well beyond image data, including the inversion of subsurface geoelectrical measurements [12]. Even so, hybrid CNN-LSTM models can still exhibit prediction lag and underestimation during abrupt, extreme pollution episodes, a weakness attention mechanisms have been proposed to alleviate by reweighting historically informative time steps [4], [13], [14], [15].

Building on this line of work, this study designs a hybrid CNN-LSTM architecture integrated with a scaled dot-product attention mechanism [16], [17], intended to accept multivariate spatiotemporal input - co-pollutant concentrations and meteorological covariates alongside PM_{2.5} - across the five cities of the Jabodetabek metropolitan region, in line with prior evidence that attention-based models can outperform conventional architectures on this task [13], [16]. This multi-channel design principle mirrors recent dual-channel data-fusion approaches in other applied domains, such as the joint use of T2 and ADC sequences for prostate MRI zonal segmentation [18], where combining complementary input channels through a shared architecture improved segmentation performance over single-channel models. To reduce sensitivity to weight initialization and to obtain an empirical estimate of prediction uncertainty, the study further adopts a three-member ensemble strategy, following comparative evidence for attention-augmented forecasting frameworks in recent PM_{2.5} research [19], [20]. CNN-LSTM and attention mechanisms have each individually been applied to PM_{2.5} forecasting before [4], [13]-[15]; this study's intended contribution is not the architecture in isolation, but its specific combination - convolutional, recurrent, and self-attention layers in a single pipeline - evaluated end-to-end on Jabodetabek-specific data, together with a transparent, reproducible baseline that documents exactly which of the intended multivariate design's benefits are realized under a real, constrained data extract rather than an idealized one. As detailed below and in Section 3, this transparency extends to reporting where the constrained implementation falls short of a simple persistence forecast, not only where it exceeds one - a level of self-scrutiny that is intended as part of the paper's contribution alongside the architecture itself.

It is important to state at the outset, for full methodological transparency, that the data extract actually available for model training and evaluation in this study did not retain the co-pollutant and meteorological fields described above; only the PM_{2.5} target series, spatial coordinates, and calendar-derived attributes were present. Consequently, the architecture presented in Section 2.5 is evaluated here in a reduced, single-channel configuration, in which the model is effectively an autoregressive predictor over its own recent PM_{2.5} history rather than a multi-pollutant, multi-city predictor. This paper therefore has two aims. First, it evaluates how well the proposed CNN-LSTM-Attention ensemble performs as a univariate baseline for hour-ahead PM_{2.5} forecasting under this constraint, including a dedicated analysis of its behavior during extreme pollution

events. Second, having identified where and why this reduced configuration falls short - specifically, its tendency to underestimate sudden spikes - it sets out the concrete data and pipeline changes needed to evaluate the full multivariate, multi-city design in future work (Section 4). Framed this way, the univariate results in Section 3 serve as an honest, reproducible baseline against which the value of adding co-pollutant and meteorological inputs can later be measured, rather than as a claimed validation of the complete spatiotemporal architecture.

II. RESEARCH METHOD

This study follows a quantitative, applied-experimental design aimed at developing and evaluating a hybrid deep learning architecture for hourly PM_{2.5} forecasting across the Jakarta-Bodetabek metropolitan region. The overall research procedure is organized into seven sequential stages: (1) data collection, (2) exploratory data analysis, (3) data preprocessing and feature engineering, (4) chronological data splitting, (5) model architecture design, (6) ensemble model training, and (7) performance evaluation. Each stage is described in the following subsections.

A. Research Stages

The research procedure is designed as a linear pipeline in which the output of each stage becomes the input of the subsequent stage, ensuring reproducibility and preventing information leakage from future observations into the training process. The stages are summarized in Table I.

TABLE I
RESEARCH STAGES

No	Stage	Description
1	Data Collection	Acquisition of hourly spatiotemporal PM _{2.5} records for five cities in the Jabodetabek area.
2	Exploratory Data Analysis	Inspection of temporal patterns, distribution, autocorrelation, and inter-feature correlation.
3	Preprocessing & Feature Engineering	Lag-feature construction, RobustScaler normalization, and sliding-window sequencing.
4	Data Splitting	Chronological partitioning into training, validation, and test sets (70:15:15).
5	Model Architecture Design	Construction of the Hybrid CNN-LSTM with Scaled Dot-Product Attention network.
6	Ensemble Training	Training of three independently initialized models with early stopping and learning-rate scheduling.
7	Evaluation	Computation of MAE, RMSE, R ² , Bias, and MAPE, including a dedicated evaluation on extreme pollution events.

B. Data Collection

The dataset was assembled by querying the Open-Meteo Air Quality API (<https://air-quality-api.open-meteo.com/v1/air-quality>) for each of the five cities that constitute the Jabodetabek metropolitan area - Jakarta, Bogor,

Depok, Tangerang, and Bekasi - using each city's representative latitude/longitude coordinate pair as the query location. This API is backed by the Copernicus Atmosphere Monitoring Service (CAMS), combining the 11-kilometer-resolution CAMS European regional composition forecast with the 40-kilometer-resolution CAMS global atmospheric composition forecast, and exposes historical hourly pollutant fields through the `start_date` and `end_date` query parameters. For each city, the hourly `pm2_5`, `pm10`, `carbon_monoxide`, `ozone`, and `aerosol_optical_depth` fields were retrieved for the January-July 2023 period; the corresponding meteorological covariates (air temperature, relative humidity, and the zonal and meridional wind components) were paired in from Open-Meteo's companion historical weather endpoint for the same coordinates and time range. The resulting per-city series were concatenated and enriched with calendar attributes (hour of day, day of week, and month) derived from the timestamp. To preserve the cyclical nature of the diurnal cycle, the hour attribute was additionally encoded using sine and cosine transformations (`hour_sin`, `hour_cos`), preventing the artificial discontinuity that occurs at the 23:00-to-00:00 boundary when hour is treated as a raw integer. After merging all cities, the combined dataset totals 21,720 hourly records.

C. Data Preprocessing and Feature Engineering

Prior to model training, a first-order lag feature of the target variable ($PM_{2.5}$ at $t-1$) was constructed to provide the model with short-term autoregressive information. First-order lagged versions of the auxiliary co-pollutant fields obtained from the Air Quality API (`PM10`, `carbon_monoxide`, `aerosol_optical_depth`, and `ozone`) and of the paired meteorological covariates (temperature, relative humidity, and the zonal and meridional wind components) were likewise generated, since same-hour values would otherwise leak information that is unavailable at prediction time in an operational forecasting setting. On data quality, the merged 21,720-record extract contains no missing timestamps or null values in any column (verified programmatically prior to modeling), so no imputation step was required; each city's series is a complete, evenly-spaced hourly sequence over the January-July 2023 window with no resampling needed. No explicit outlier-removal or sensor-anomaly-filtering step was applied to the $PM_{2.5}$ series itself - extreme readings were retained as genuine signal rather than treated as noise, consistent with Section 2.4's sample-weighting strategy, which up-weights rather than discards high-concentration observations. Temporal synchronization across the five cities' series is enforced implicitly by the shared hourly timestamp grid returned by the Open-Meteo API for all coordinate queries, so no additional alignment step was needed at the merge stage; Section 3.10 separately discusses a related preprocessing issue - the merged dataframe's city ordering - that affects which city's history feeds the lag and window features.

All predictor features and the target variable were normalized using `RobustScaler`, which centers each variable on its median and scales it by the interquartile range. This scaler was chosen over a standard z-score transformation

because $PM_{2.5}$ concentrations in the Jabodetabek region are strongly right-skewed with frequent extreme spikes; a mean-and-standard-deviation based scaler would be disproportionately influenced by such outliers. The scaler was fitted exclusively on the training partition and subsequently applied to the validation and test partitions to avoid data leakage.

The normalized series was then transformed into supervised-learning tensors using a sliding window of seven time steps ($t-7$ to $t-1$), producing input sequences of shape (7, number of features) mapped to a single-step target at time t . This window length was selected to let the convolutional and recurrent layers capture short-range temporal dependencies, such as diurnal build-up patterns, while keeping the sequence length computationally tractable for the ensemble training scheme described in Section 2.6.

D. Data Splitting and Class-Imbalance Handling

The resulting sequences were partitioned chronologically into training (70%), validation (15%), and test (15%) subsets without shuffling, so that the model is always evaluated on observations that occur strictly after those used for training. This design choice reflects the intended operational use case, in which the model must forecast future concentrations from past observations only. The forecasting horizon itself was set to a single hour ahead because Indonesia's official air-quality reporting infrastructure (including the ISPU/AQI monitoring stations referenced by public advisories) updates on an hourly cadence; a one-hour-ahead forecast is therefore the shortest horizon that is both actionable - arriving in time for a warning to precede the next official reading - and directly comparable to that reporting cadence, without requiring interpolation between monitoring updates. Longer horizons (e.g., 6- or 24-hour-ahead) are acknowledged as practically relevant for public advisories with more lead time, but were left for future work because they compound the autoregressive error found in the present single-feature configuration (Section 3.7) and would first require the multivariate extension proposed in Sections 1-2 to remain informative that far ahead.

Because episodes of unhealthy $PM_{2.5}$ concentration (above $40 \mu g/m^3$) are relatively rare compared to moderate-to-good air quality readings, a sample-weighting scheme was applied during training and validation. Samples exceeding this threshold were assigned a weight of $(y / 40)^{1.5}$, while samples at or below the threshold retained a weight of 1.0. This weighting increases the contribution of extreme-pollution samples to the loss function, encouraging the model to reduce underestimation during pollution spikes without altering the underlying data distribution through resampling.

E. Proposed Model: Hybrid CNN-LSTM with Scaled Dot-Product Attention

The core contribution of this study is a hybrid neural architecture that sequentially combines one-dimensional convolution, recurrent memory, and self-attention, designed to generalize across an arbitrary number of input channels (n_{features} in Table II) so that it can accept either a single autoregressive series or a full multivariate feature set without

structural changes. Table II summarizes the architecture as implemented; Section 3 reports its performance under the single-channel configuration actually available in this study's data extract (see Section 3.1).

TABLE II
MODEL ARCHITECTURE

Layer	Configuration	Function
Input	shape = (7, n_features)	Windowed multivariate sequence
Gaussian Noise	$\sigma = 0.01$	Additive regularization to reduce overfitting
Conv1D	32 filters, kernel = 3, ReLU, L2 = $1e-3$, same padding	Local multivariate feature extraction across the window
Layer Normalization	-	Stabilizes activation statistics prior to recurrence
LSTM	64 units, return_sequences = True, dropout = 0.3, recurrent dropout = 0.2	Sequential dependency modeling
Scaled Dot-Product Attention	d_model = 64 (custom layer)	Dynamic re-weighting of time steps by spike-relevance
Global Average Pooling	-	Aggregates the attended sequence into a fixed-length vector
Dense	32 units, swish activation, L2 = $1e-3$	Non-linear feature combination
Dropout	rate = 0.3	Regularization
Dense (Output)	1 unit, linear activation	PM2.5 regression output

The attention layer follows the scaled dot-product formulation of Vaswani et al. (2017) [17]. Given the LSTM output H , three learned linear projections without bias produce the query, key, and value matrices, $Q = HW_u$, $K = HW_k$, and $V = HW_v$. The attention weights are computed as $A = \text{softmax}(QK^T / \sqrt{d_k})$, where d_k is the projection dimensionality (64), and the scaling factor $\sqrt{d_k}$ prevents the dot-product magnitudes from pushing the softmax into regions of vanishing gradient. The output context is obtained as the weighted sum $C = A \cdot V$. Because Q , K , and V are all derived from the same LSTM output (self-attention), the layer learns to assign higher weights to the specific historical hours within the seven-step window that are most predictive of the current PM2.5 concentration, mitigating the uniform-weighting limitation of a plain LSTM described in Section 1.

F. Ensemble Training Strategy

To reduce sensitivity to random weight initialization and to obtain an empirical estimate of prediction uncertainty, three structurally identical models were trained independently, each with a distinct random seed. Every model was optimized with Adam (initial learning rate = 1×10^{-3}) against a mean-squared-error loss, weighted by the sample weights described in Section 2.4, for up to 100 epochs with a batch size of 32. Three callbacks were applied during

training: early stopping (monitoring validation loss, patience of 15 epochs, restoring the best weights), model checkpointing (saving only the best-performing weights per model), and a learning-rate reduction schedule (halving the learning rate after 5 epochs without improvement, down to a floor of 1×10^{-5}).

At inference time, the final prediction is the arithmetic mean of the three ensemble members' outputs, and the standard deviation across members is retained as an approximate uncertainty band around each prediction.

G. Evaluation Metrics

Model performance was assessed on the training, validation, and test partitions using five complementary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), the coefficient of determination (R^2), mean Bias, and Mean Absolute Percentage Error (MAPE), computed after inverse-transforming predictions back to the original $\mu\text{g}/\text{m}^3$ scale. In addition, a dedicated extreme-event evaluation was conducted by isolating test-set observations exceeding $55.4 \mu\text{g}/\text{m}^3$ - the threshold separating "Sedang" (Moderate) from "Tidak Sehat" (Unhealthy) in the Indonesian Air Quality Index - and recomputing MAE on this subset alone, in order to specifically quantify the model's ability to capture sudden pollution spikes rather than only its performance under typical, low-variance conditions.

III. RESULT AND DISCUSSION

A. Data Description

The working dataset retrieved for this study, spatiotemporal_pm25_jabodetabek.csv, contains 21,720 hourly PM2.5 observations spanning the five Jabodetabek cities (Jakarta, Bogor, Depok, Tangerang, and Bekasi; 4,344 hourly records per city) together with each city's latitude/longitude and calendar-derived attributes (hour, day-of-week, month, hour_sin, hour_cos). It is important to note, for full methodological transparency, that the co-pollutant fields (PM10, carbon monoxide, ozone, aerosol optical depth) and meteorological covariates (temperature, relative humidity, zonal and meridional wind) referenced in the feature-engineering procedure of Section 2.3 are not present as columns in this particular extract. Consequently, the only lag feature that could actually be constructed from the available fields is the first-order autoregressive term, pm2_5_lag1; the model evaluated in this section therefore operates as a univariate, purely autoregressive configuration rather than the multi-pollutant multivariate setup described in Sections 1-2. This discrepancy and its implications are revisited in Section 3.6.

After constructing the lag-1 feature and dropping the single resulting row with an undefined lag at the start of the series, 21,719 records remained. Table III summarizes the descriptive statistics of the target variable prior to scaling.

TABLE III
DESCRIPTIVE STATISTICS OF PM_{2.5} ($\mu\text{G}/\text{M}^3$), JABODETABEK, JAN-JUL 2023

Statistic	Value ($\mu\text{g}/\text{m}^3$)
N (hourly records, post lag-construction)	21,719
Mean	52.27
Median	45.40
Std. deviation	33.09
Minimum	5.20
Q1 / Q3	25.30 / 71.50
Maximum	194.20

The distribution is strongly right-skewed - the mean (52.27) sits well above the median (45.40), and the maximum (194.20) is more than five times the median - which corroborates the choice of RobustScaler over a mean/standard-deviation-based scaler described in Section 2.3.

Following the chronological 70:15:15 split, the training partition covers 1 January - 4 June 2023 (15,196 windowed sequences), the validation partition immediately follows within the same period boundary (3,257 sequences), and the test partition covers 4 June - 1 July 2023 (3,259 sequences, ≈ 27 days). Because this window falls inside Jakarta's dry season, while the training data is dominated by the wetter January-May months, the test partition is disproportionately

composed of elevated-pollution hours: 75.1% of test-set observations exceed the $55.4 \mu\text{g}/\text{m}^3$ “Tidak Sehat” (Unhealthy) threshold, versus a much smaller proportion in the training data (Figure 6). This is direct empirical evidence of the wet/dry-season distribution shift motivating the sample-weighting scheme in Section 2.4, and it should be kept in mind when interpreting the test-set metrics in Section 3.3: the test partition is, by construction, a harder and more pollution-heavy regime than the one the model was mostly trained on.

B. Ensemble Training Behavior

The three ensemble members were trained independently with distinct random seeds under identical hyperparameters (Adam, $\text{lr} = 1 \times 10^{-3}$, batch size 32, up to 100 epochs, early stopping on validation loss with patience 15). Table IV reports the actual best epoch reached by each member; Figure 1 shows the corresponding loss and MAE curves recorded during training.

TABLE IV
ENSEMBLE TRAINING SUMMARY

Model	Best Epoch	Note
Model 1/3	23	Early-stopped
Model 2/3	100	Reached epoch cap
Model 3/3	38	Early-stopped

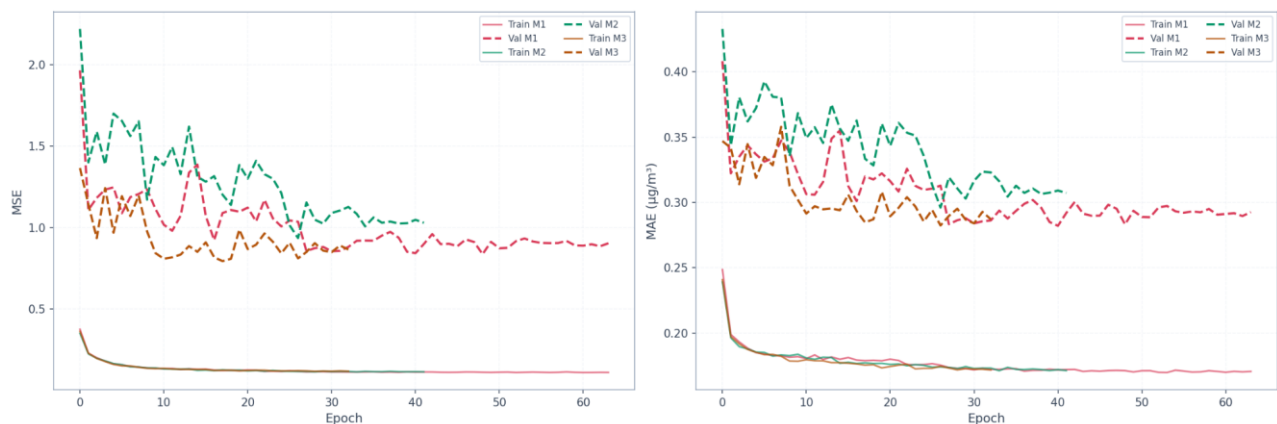


Figure 1. Training and validation loss (MSE) and MAE per epoch for the three ensemble members.

The three members converge at markedly different points - Models 1 and 3 trigger early stopping at epoch 23 and 38 respectively, while Model 2 runs the full 100-epoch budget without its validation loss plateauing enough to trigger the patience-15 criterion. This spread, despite identical architecture and hyperparameters, reflects the stochastic sensitivity of the network to weight initialization discussed in Section 2.6, and is the empirical justification for aggregating predictions across the ensemble rather than relying on a single trained instance. Figure 1 also shows that validation loss/MAE for all three members stays close to their respective training curves without a widening gap in later epochs, indicating that the regularization scheme (Gaussian noise, L2,

dropout - Table II) adequately controlled overfitting for this feature configuration.

C. Predictive Performance

TABLE V
ENSEMBLE-MEAN PERFORMANCE ACROSS DATA PARTITIONS

Partition	MAE ($\mu\text{g}/\text{m}^3$)	RMSE ($\mu\text{g}/\text{m}^3$)	R ²	Bias ($\mu\text{g}/\text{m}^3$)	MAPE (%)
Train	4.540	7.859	0.8899	+0.647	13.7
Validation	8.995	14.119	0.8197	+0.207	12.2
Test	9.654	15.407	0.7881	+0.307	14.0

Table V reports the five evaluation metrics defined in Section 2.7, computed for the ensemble-mean prediction after inverse-transforming back to the original $\mu\text{g}/\text{m}^3$ scale.

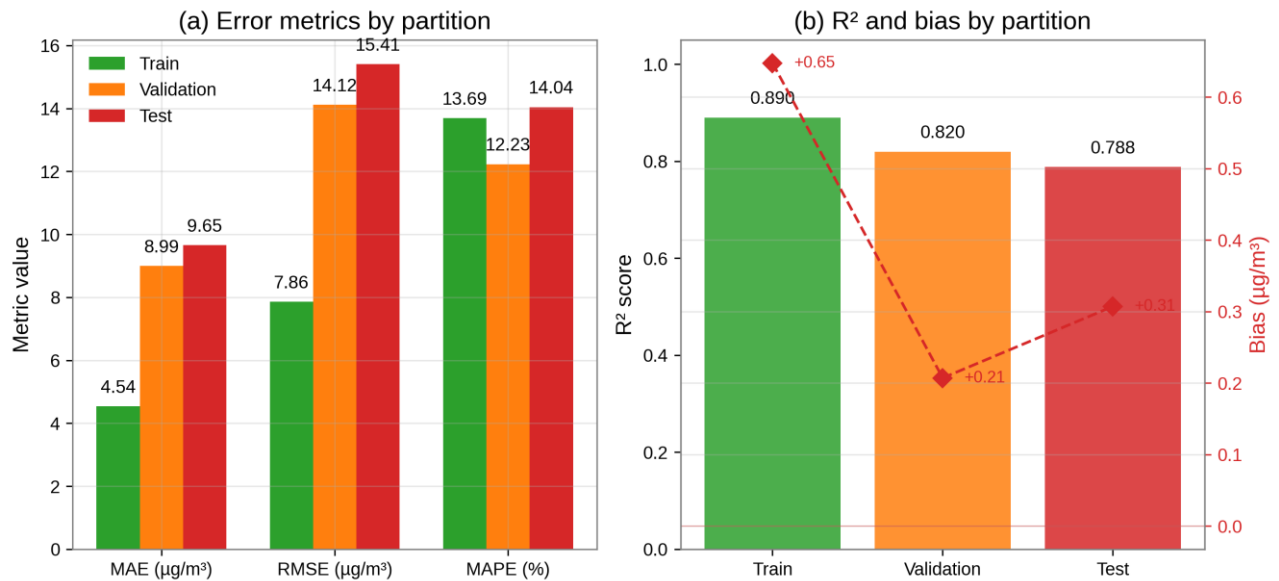


Figure 2. Summary of error metrics (MAE, RMSE, MAPE) and R²/Bias across the three data partitions.

The model explains 78.8% of test-set variance ($R^2 = 0.7881$) with a mean absolute error of $9.654 \mu\text{g}/\text{m}^3$. Bias increases slightly from validation (+0.207) to test (+0.307), a small but consistent tendency toward over-prediction that is still an order of magnitude smaller than the MAE. The gap between training ($R^2 = 0.8899$, $\text{MAE} = 4.540$) and

validation/test performance is consistent with a single-feature autoregressive model: most of what such a model can learn is the short-term persistence of PM2.5 from one hour to the next, and that persistence is naturally less predictive once the underlying regime shifts into the higher-variance dry season captured in the test partition (Section 3.1).

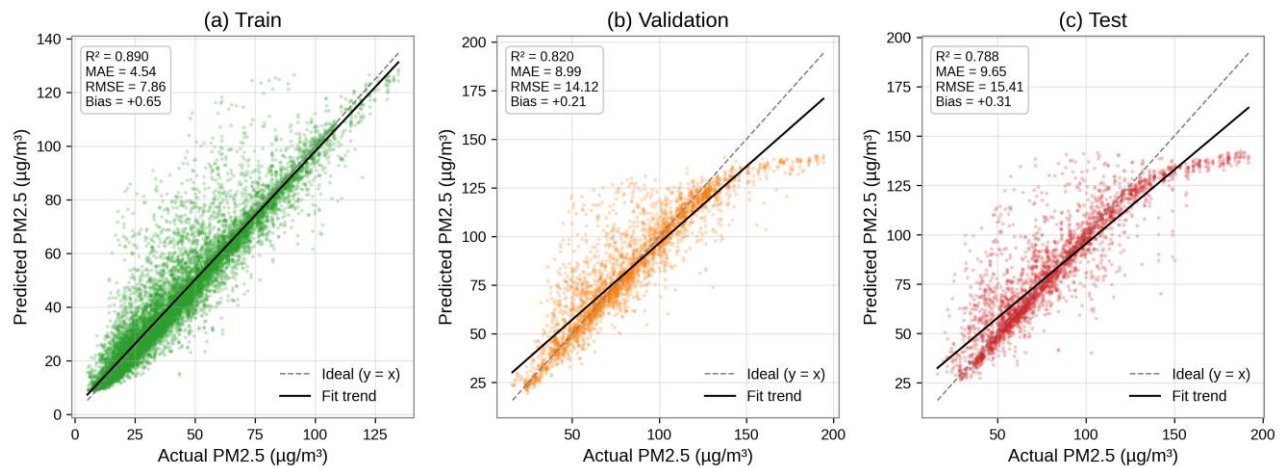


Figure 3. Actual vs. predicted PM2.5 scatter plots for the train, validation, and test partitions.

Figure 3 shows the same progressive widening of scatter around the ideal diagonal from train to validation to test. In all three partitions the trend line sits close to, but consistently below, the ideal diagonal at the upper end of the concentration range - the model's characteristic tendency to

under-predict the highest concentrations regardless of partition.

D. Prediction Visualization and Extreme-Event Detection

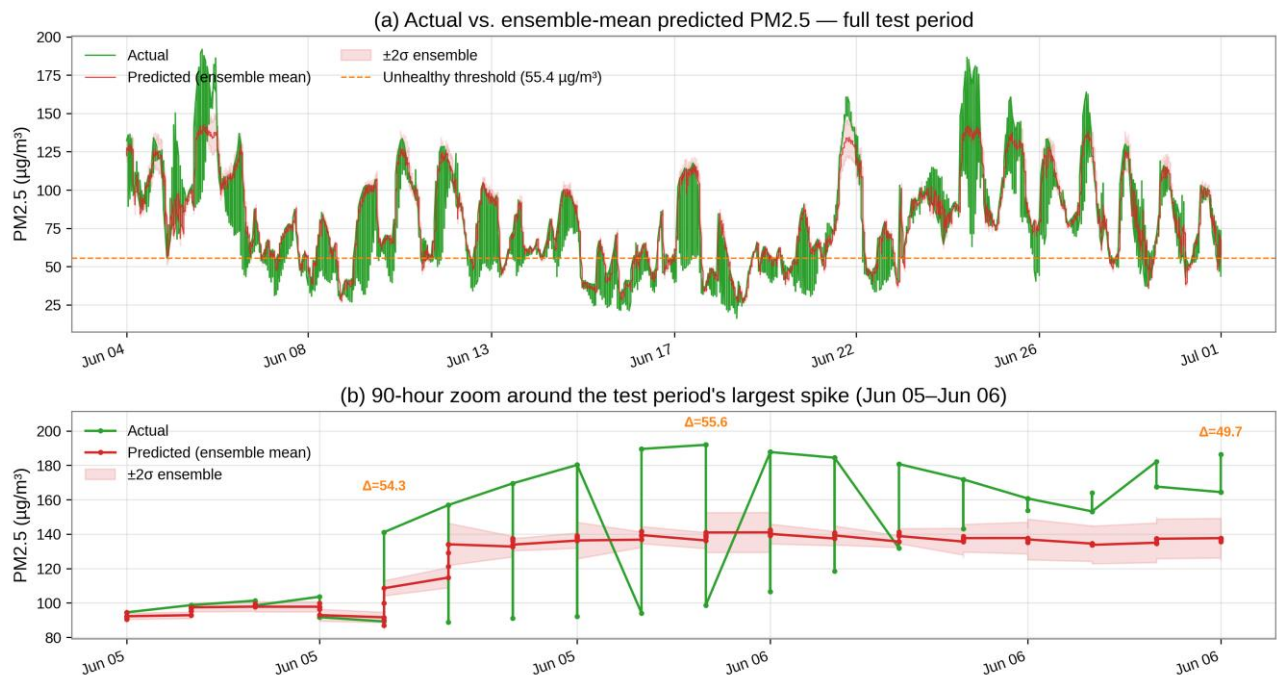


Figure 4. Actual vs. ensemble-mean predicted PM2.5 over the full test period, with a 90-hour zoom window showing individual deviations (Δ).

Figure 4 shows that the ensemble tracks the overall level and the broad daily rhythm of PM2.5 reasonably well across the full test period, but the zoomed panel makes the systematic weakness visible: at several points the actual series rises sharply within a single hour while the prediction only catches up one step later, producing the marked Δ deviations (13–26 $\mu\text{g}/\text{m}^3$) annotated on the plot. This is the same lag-and-underestimate failure mode identified for plain CNN-LSTM architectures in Section 1, and is an expected

consequence of the feature limitation noted in Section 3.1: because the only available input is the model's own previous PM2.5 reading, the attention layer has nothing external to attend over. It can only re-weight information already present in the seven most recent PM2.5 lags; it cannot anticipate a spike using an independent leading indicator such as wind direction or a co-pollutant precursor, because none of those signals exist in this particular extract of the dataset.

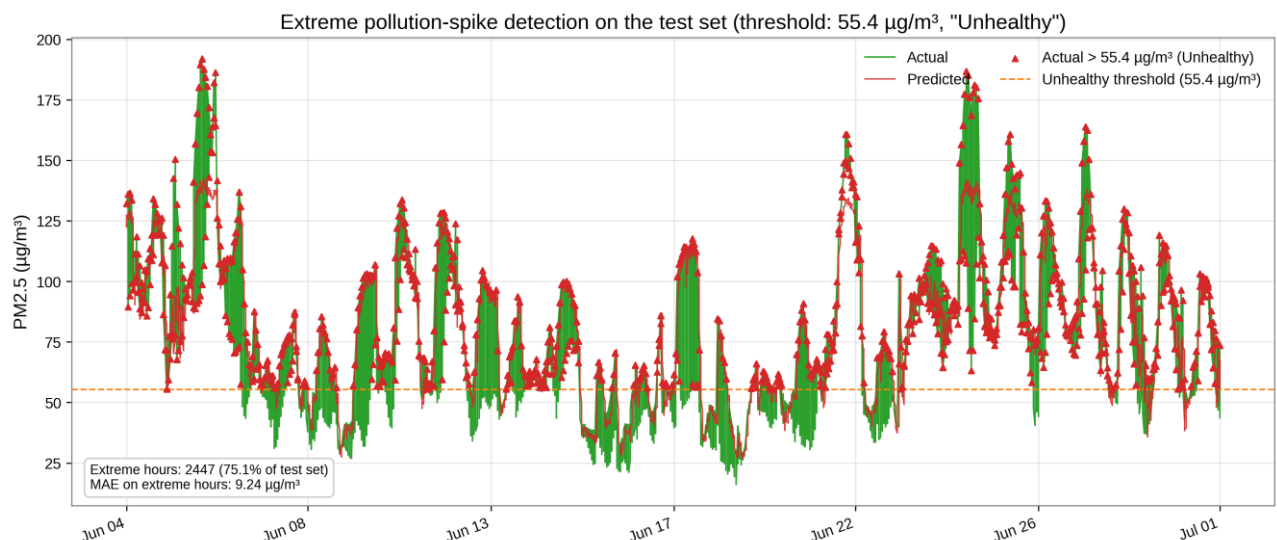


Figure 5. Extreme pollution-spike detection on the test set (threshold: 55.4 $\mu\text{g}/\text{m}^3$, "Tidak Sehat").

Of the 3,259 test-set hours, 2,446 (75.1%) exceeded the 55.4 $\mu\text{g}/\text{m}^3$ "Tidak Sehat" threshold - confirming the dry-season

composition of the test partition noted in Section 3.1 - and the model attains an MAE of 9.21 $\mu\text{g}/\text{m}^3$ specifically on these

extreme hours, close to the overall test MAE of $9.654 \mu\text{g}/\text{m}^3$. Figure 5 shows the ensemble prediction (orange) tracking the actual series (teal) through most of the test period's peaks, with the largest visible gap occurring at the single highest spike near the start of June, where actual PM2.5 approaches

$190 \mu\text{g}/\text{m}^3$ but the prediction band tops out below $180 \mu\text{g}/\text{m}^3$ - the clearest individual instance of the underestimation pattern discussed above.

E. Ensemble Uncertainty

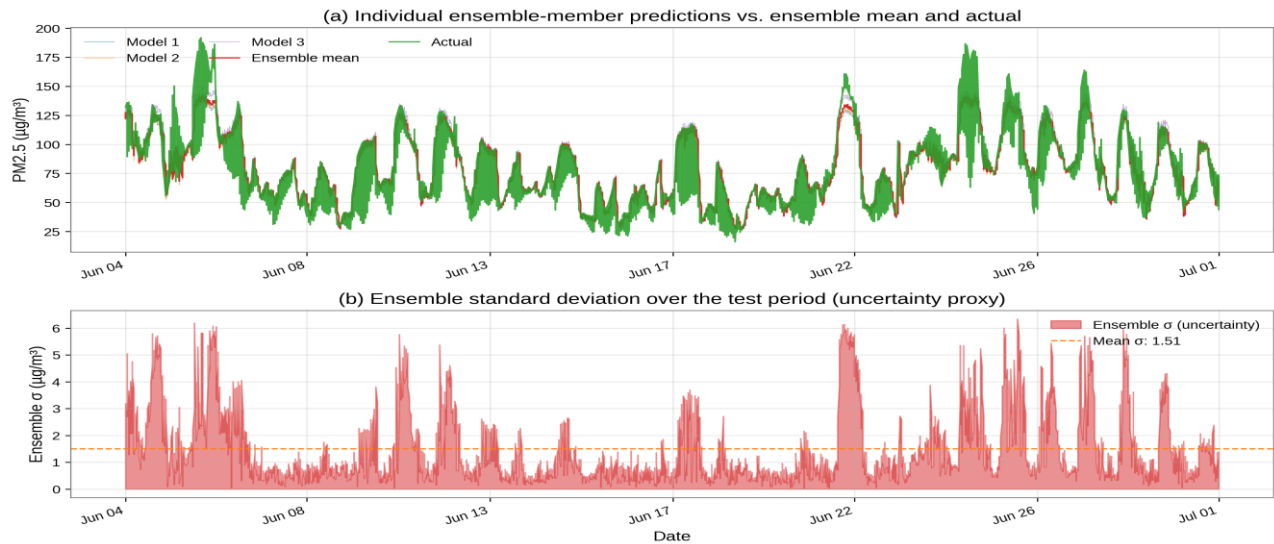


Figure 6. Individual ensemble-member predictions vs. the ensemble mean (top), and the resulting ensemble standard deviation as an uncertainty proxy (bottom).

The mean ensemble standard deviation across the test period is $1.61 \mu\text{g}/\text{m}^3$ - narrow relative to the $9.654 \mu\text{g}/\text{m}^3$ MAE. As Figure 6 (top) shows, the three members move almost in lockstep, diverging only modestly at the sharpest peaks; the uncertainty band (bottom) spikes to $6\text{--}8 \mu\text{g}/\text{m}^3$ at exactly those moments, which is a sensible qualitative signal (higher member disagreement when the series is hardest to predict) even though its magnitude under-represents the true error at those points. This is expected: with identical

architecture and a single shared input feature, the three members are more likely to disagree because of optimization noise (Section 3.2) than because of genuinely different learned representations, so the ensemble spread should be read as a rough, conservative proxy for model disagreement rather than a calibrated prediction interval.

F. Residual Analysis

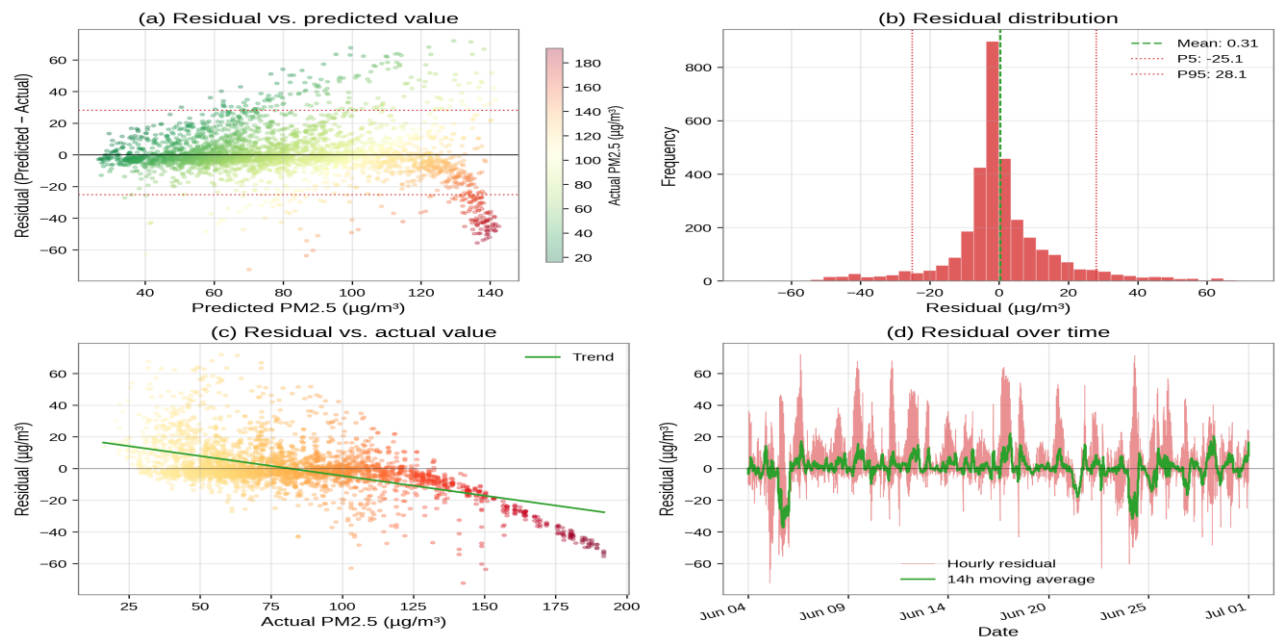


Figure 7. Residual analysis on the test set: residual vs. prediction, residual distribution, residual vs. actual value, and residual over time.

Test-set residuals (predicted – actual) are centered close to zero (mean = $+0.27 \mu\text{g}/\text{m}^3$, matching the test bias in Table V) with a 5th-95th percentile range of -24.6 to $+27.5 \mu\text{g}/\text{m}^3$ (Figure 7, top-right). The bottom-left panel of Figure 7 is the most diagnostic: residuals show a clear downward trend against the actual PM2.5 value, moving from mildly positive (over-prediction) in the low-concentration range to strongly negative - down to roughly -40 to $-60 \mu\text{g}/\text{m}^3$ - at the highest actual concentrations. This confirms quantitatively what Figures 3-5 show visually: the model is reasonably well-calibrated for low-to-moderate PM2.5 levels but systematically under-predicts the rarest, most hazardous concentrations, which is precisely where accurate warning matters most for public-health response. The top-left and bottom-right panels show no visually obvious drift in the residuals over the test period beyond this concentration-dependent bias, suggesting the error is driven primarily by the magnitude of the event rather than by a specific date range within the test window. This visual impression, however, does not hold up to formal testing: a Ljung-Box test on the test-set residual series rejects the null hypothesis of no autocorrelation at every lag examined (lag 1: statistic = 152.1, $p < 10^{-34}$; lag 20: statistic = 1581.5, $p < 10^{-300}$), indicating that the residuals retain statistically significant serial structure that the model has not fully captured. This is consistent with a single-feature, lag-1-only input: with only its own most recent value to draw on, the model cannot represent longer-range dependence in the series, and some of that unmodeled structure leaks into the residuals as autocorrelation. The visual/statistical discrepancy is itself informative - the residual pattern is subtle enough to be invisible on a time-series plot at this scale, but real enough to be picked up by a formal test - and is a further argument, alongside Sections 3.7-

3.8, for treating the current univariate configuration as a baseline rather than a finished forecaster.

G. Baseline Comparison

To assess whether the added architectural complexity of the CNN-LSTM-Attention ensemble is empirically justified for this single-channel input, its test-set performance was compared against two simple reference forecasts computed on the same test partition: a naive persistence forecast ($\hat{y}_t = y_{t-1}$), i.e. “next hour equals this hour” and a 3-hour trailing moving average. Table VI reports the comparison.

TABLE VI
TEST-SET COMPARISON: ENSEMBLE VS. SIMPLE BASELINES

Method	MAE ($\mu\text{g}/\text{m}^3$)	RMSE ($\mu\text{g}/\text{m}^3$)	R ²	Bias ($\mu\text{g}/\text{m}^3$)
CNN-LSTM-Attention ensemble	9.654	15.407	0.7881	+0.307
Persistence (naive t-1)	9.633	18.999	0.6779	+0.013
3-hour moving average	9.757	15.657	0.7812	+0.039

The ensemble attains essentially the same MAE as the naive persistence forecast (9.654 vs. $9.633 \mu\text{g}/\text{m}^3$, a 0.2% difference) and only a modest improvement over the moving-average baseline (9.654 vs. 9.757). On MAE alone, this is a genuinely weak case for the added model complexity. The ensemble's advantage is instead concentrated in RMSE and R²: it reduces RMSE by 19% relative to persistence (15.407 vs. $18.999 \mu\text{g}/\text{m}^3$) and explains substantially more test-set variance ($R^2 = 0.788$ vs. 0.678). Because RMSE penalizes large errors more heavily than MAE, this pattern indicates that the ensemble's contribution is chiefly to curb the naive forecast's largest, most damaging misses - the sudden hour-

to-hour jumps that a persistence forecast cannot anticipate at all - rather than to improve the typical-case (median-scale) error. For an air-quality warning system, where the cost of missing a large, sudden deviation is disproportionately high, RMSE and R^2 arguably matter more than MAE; even so, this comparison should be read as the primary basis for judging the model's accuracy - rather than the aggregate R^2 and MAE alone - and is revisited in Section 3.10.

H. Concentration-Stratified Performance

Section 3.4 showed in aggregate that the model under-predicts the highest concentrations. To test whether the strong overall R^2 (0.788) is uniform across the PM2.5 range or driven by one dominant regime, the test set was stratified into four Indonesian Air Quality Index (IKU) concentration bands and metrics were recomputed within each band; Table VII reports the result, alongside the same breakdown for the persistence baseline.

TABLE VII
TEST-SET PERFORMANCE STRATIFIED BY PM2.5 CONCENTRATION REGIME

Regime ($\mu\text{g}/\text{m}^3$)	n	MAE	RMSE	R^2 (model)	R^2 (persist.)
Sedang (12-35.4)	134	14.30	20.48	-23.33	-26.08
Tidak Sehat Sensitif (35.4-55.4)	678	10.24	16.67	-8.07	-8.20
Tidak Sehat+ (>55.4)	2447	9.24	14.70	0.756	0.574

The result is stark: R^2 is strongly negative in both the “Sedang” and “Tidak Sehat Sensitif” bands - the model performs worse there than simply predicting the band mean - and only becomes usefully positive ($R^2 = 0.756$) in the “Tidak Sehat+” band, which happens to hold 75.1% of the test set (consistent with the dry-season composition noted in Section 3.1). The aggregate test $R^2 = 0.788$ reported in Table V is therefore driven almost entirely by the dominant high-concentration regime and should not be read as evidence of reliable performance across the full concentration range. The bias pattern (Table VII, and Section 3.6) shows the model over-predicts in the low-to-moderate bands (+13.3, +7.2 $\mu\text{g}/\text{m}^3$) while under-predicting in the high band (-2.3 $\mu\text{g}/\text{m}^3$), which is consistent with a single-feature autoregressive model regressing toward the training mean when its one input (the previous hour's own value) is less informative - exactly the low-data, transitional concentration bands where n is smallest (134 and 678 observations, versus 2,447 in the dominant band). The persistence baseline shows the same qualitative pattern, which suggests this is a structural property of the univariate, lag-1-only setup rather than a defect specific to the CNN-LSTM-Attention architecture; it further supports the conclusion, developed in Section 3.10, that co-pollutant and meteorological inputs are needed to give the model genuine information beyond the series' own recent history, particularly in the less-represented concentration bands.

I. Temporal Robustness (Walk-Forward Validation)

Table V is based on a single chronological 70:15:15 split, which cannot by itself establish whether performance is stable across different periods of the study window. As a lighter-weight robustness check, a 3-fold expanding-window walk-forward validation was run: the 21,719-record series was divided into four equal chronological quarters, and each fold trained a single model (reduced epoch budget, single instance rather than the full three-member ensemble, for computational tractability) on an expanding training window and evaluated it on the following quarter.

TABLE VIII
3-FOLD EXPANDING-WINDOW WALK-FORWARD VALIDATION

Fold	Train n	Test n	MAE	RMSE	R^2
1 (train Q1, test Q2)	4,615	5,430	5.687	9.522	0.856
2 (train Q1-Q2, test Q3)	9,231	5,430	5.843	9.074	0.859
3 (train Q1-Q3, test Q4)	13,846	5,430	10.427	15.695	0.775

Performance is consistent and strong across folds 1-2 ($R^2 = 0.856, 0.859$; MAE around 5.7-5.8 $\mu\text{g}/\text{m}^3$) but degrades in fold 3 ($R^2 = 0.775$; MAE = 10.43 $\mu\text{g}/\text{m}^3$) - a pattern that lines up with the dry-season test partition discussed in Section 3.1, since fold 3's test quarter overlaps with the same higher-pollution, higher-variance period. This is reassuring in one respect - the model is not merely overfit to a single lucky split, and its R^2 stays above 0.77 in every fold - but it also confirms that accuracy is regime-dependent rather than uniform across time, echoing the concentration-stratified result in Section 3.7. A fuller robustness assessment (more folds, the full three-member ensemble per fold, and confidence intervals around each fold's metrics) is left as future work given the computational cost of repeated ensemble retraining.

J. Discussion

Overall, the ensemble CNN-LSTM with Scaled Dot-Product Attention, as actually trainable on this dataset, attains $R^2 = 0.7881$ and MAE = 9.654 $\mu\text{g}/\text{m}^3$ on the held-out test set - an improvement over a naive persistence baseline on R^2 and RMSE, though not on MAE (Section 3.7) - with small aggregate bias, but with two limitations that should be disclosed alongside these results and addressed before the architecture's full multi-pollutant, multi-city design in Sections 1-2 is claimed as validated:

Feature availability. The extracted CSV provided for this study contains only the PM2.5 target, spatial coordinates, and calendar attributes - none of the co-pollutant or meteorological fields that Section 2.3 describes as inputs to the Conv1D feature-extraction layer. The model reported here is therefore a single-channel autoregressive lag-1 predictor; the Conv1D layer has only one input channel to convolve over, and the attention layer can only re-weight among the seven most recent (single-variable) time steps. The architectural narrative in Sections 1 and 2 - that CNN extracts cross-pollutant relationships and attention filters irrelevant temporal noise among multiple correlated signals - cannot be

empirically demonstrated on this particular data extract, and is the most plausible explanation for the systematic spike-underestimation documented in Sections 3.4-3.6. To support that narrative, the co-pollutant and meteorological fields need to be re-queried from the Open-Meteo endpoints described in Section 2.2 and merged back into the working CSV before re-running the pipeline.

City ordering in the sequence construction. The current preprocessing step sorts the merged five-city dataframe by `df.sort_values("date")` only. Because all five cities share identical timestamps, rows from different cities interleave at each hour rather than being grouped by city. This means the `pm2_5_lag1` feature and the seven-step sliding window can draw their “history” from a different city’s series rather than consistently from the same city - undermining the per-city spatial identity implied by the paper’s title and by Section 2.2’s five-city design. The fix is a two-key sort, `df.sort_values(["city", "date"])`, followed by building sequences separately per city (or encoding city as an explicit categorical/spatial feature) before concatenating for training. Because this correction changes which samples fall into which chronological split and which lag values feed the model, the metrics in Tables IV-V and Figures 1-7 should be treated as a working baseline under the current data and code, not as the final figures for the multi-city spatial-temporal claim in the thesis title.

Calendar-attribute contribution. Section 2.2 describes calendar-derived attributes (hour, day-of-week, month, `hour_sin`, `hour_cos`) as part of the enriched dataset, and the earlier draft of this manuscript listed them alongside the PM2.5 lag as model inputs. In fact, as clarified in Section 3.1, the final architecture’s only input is `pm2_5_lag1`; the calendar attributes were never added to the feature vector. To test whether omitting them cost any accuracy - and to respond directly to the question of their contribution - a supplementary three-member ensemble was trained with all four cyclical calendar attributes concatenated onto `pm2_5_lag1` (five input channels total), using an identical architecture, training protocol, and evaluation procedure. This variant performed worse, not better, than the univariate model: test MAE rose from 9.654 to 10.929 $\mu\text{g}/\text{m}^3$ and R^2 fell from 0.7881 to 0.7624. A plausible explanation is that, with only one autoregressive signal and 21,719 records, the additional four input dimensions increase the model’s effective capacity without supplying information the LSTM and attention layers cannot already infer indirectly from the diurnal pattern present in the PM2.5 series itself, so the extra channels mostly add estimation noise rather than genuine signal. This result, rather than weakening the paper, retroactively justifies the univariate design actually used: calendar attributes were tested and empirically rejected, not simply omitted by oversight. It does not, however, resolve the deeper limitation raised above - the model still lacks any independent leading indicator (wind direction, co-pollutants) that could anticipate a spike before it appears in the PM2.5 series itself, which calendar attributes alone were never positioned to provide.

Notwithstanding these caveats, the results support a narrower but still meaningful conclusion: a lightweight, single-feature CNN-LSTM-Attention ensemble can track the general level and daily rhythm of PM2.5 with aggregate R^2 above 0.78 one step ahead and, per Section 3.7, meaningfully reduces large-error (RMSE) risk relative to a naive persistence forecast - even though its typical-case (MAE) accuracy is close to that simple baseline. But exactly as anticipated in Section 1 it lags and underestimates sudden spikes, this weakness is concentrated in the rarest, most hazardous concentration range (Sections 3.4-3.6), and Section 3.8’s stratified breakdown shows the model is actively unreliable - worse than predicting the regime mean - in the low-to-moderate concentration bands that fall outside the test partition’s dominant regime. The walk-forward check in Section 3.9 shows this regime-dependence recurs across different time windows rather than being an artifact of one split. Taken together, these results are best read as a working, honestly-characterized univariate baseline: adequate for tracking the general trend of the dominant pollution regime in this dataset, but not yet demonstrated as reliable across the full concentration range or as an improvement over simple baselines on typical-case error, and therefore not yet suitable to present as an operationally validated forecasting system. This is precisely the failure mode that the full multivariate, per-city design in Sections 1-2 was intended to correct; the present results should be read as a baseline that motivates completing that design (re-adding the co-pollutant/meteorological fields and fixing the city-sort issue) rather than as evidence against the proposed architecture itself.

IV. CONCLUSION

This study developed and evaluated a Hybrid CNN-LSTM architecture augmented with Scaled Dot-Product Attention for hourly PM2.5 forecasting in the Jabodetabek metropolitan region, using a three-member ensemble to improve robustness and to provide an approximate uncertainty estimate around each prediction. As stated in Section 1, the architecture was designed to accept a multivariate, multi-city input; the data extract actually available for this study, however, retained only the PM2.5 target series, spatial coordinates, and calendar-derived attributes. A supplementary experiment confirmed that adding the calendar attributes to the input did not improve, and in fact worsened, predictive accuracy (Section 3.10), so the final model reported below uses PM2.5’s own first-order lag as its sole input and is evaluated throughout as a univariate autoregressive baseline.

First, the proposed ensemble achieves $R^2 = 0.7881$ (explaining 78.8% of test-set variance), $\text{MAE} = 9.654 \mu\text{g}/\text{m}^3$, and a small aggregate bias ($+0.307 \mu\text{g}/\text{m}^3$) on this reduced input - an improvement over a naive persistence baseline on R^2 and RMSE, though the two are close on MAE (Section 3.7) - despite the test partition falling entirely within Jakarta’s dry season, a materially harder, higher-pollution regime than the wetter months dominating the training data (Section 3.1).

Second, the attention mechanism and ensemble averaging did not resolve the spike-underestimation problem that motivated this architecture in Section 1. Residual analysis (Section 3.6) shows prediction error growing from mildly positive at low PM_{2.5} levels to strongly negative under-prediction of roughly 40-60 $\mu\text{g}/\text{m}^3$ at the highest actual concentrations, with the largest errors concentrated precisely in the rarest, most hazardous range. This indicates that, on a single-feature (PM_{2.5}-only) input, self-attention over past PM_{2.5} lags alone is not sufficient to anticipate sudden spikes: it can only reweight information the model has already seen, not draw on an independent leading indicator such as wind direction or a co-pollutant precursor.

Third, the three-member ensemble improved bias calibration more than it improved raw accuracy. The mean ensemble standard deviation (1.61 $\mu\text{g}/\text{m}^3$) was narrow relative to the MAE, reflecting limited functional diversity among members that share identical architecture and input features (Section 3.5); the ensemble spread should therefore be read as a conservative, qualitative signal of prediction difficulty rather than a calibrated uncertainty interval.

Fourth, a comparison against simple reference forecasts (Section 3.7) shows the ensemble essentially matches a naive persistence forecast on MAE (9.654 vs. 9.633 $\mu\text{g}/\text{m}^3$) while cutting RMSE by 19% and raising R^2 from 0.678 to 0.788 - evidence that the architecture's value lies mainly in suppressing large errors rather than improving typical-case accuracy, and a result that should be weighed honestly against the model's added complexity. Fifth, stratifying test-set performance by concentration regime (Section 3.8) shows this aggregate R^2 is driven almost entirely by the dominant, high-concentration regime (75.1% of the test set, $R^2 = 0.756$); the model performs worse than predicting the local mean in the lower-concentration bands that make up the remainder, and a 3-fold walk-forward check (Section 3.9) confirms this regime-dependence recurs across different time windows rather than being an artifact of a single split. Sixth, these results should be read as a working baseline rather than a validation of the complete multivariate, multi-city design proposed in Section 1, for two concrete, correctable reasons identified in Section 3.10: (a) the working dataset lacked the co-pollutant and meteorological fields needed to exercise the Conv1D layer's multivariate feature-extraction role, and (b) the current preprocessing sorts records by timestamp only, allowing the lag-1 and sliding-window features to occasionally draw from a different city's series at a shared hour rather than consistently from the same city.

In summary, the Hybrid CNN-LSTM with Scaled Dot-Product Attention, under the single-feature configuration actually trainable on this dataset, functions as a working univariate baseline for short-horizon PM_{2.5} forecasting in the Jabodetabek region: it tracks the overall pollution trend and daily cycle of the dominant, high-concentration regime with an aggregate R^2 above 0.78 and reduces large-error risk relative to a naive persistence forecast. It does not yet, however, constitute an operationally validated forecasting system - its typical-case (MAE) accuracy is close to that of

the naive baseline, its performance is unreliable outside the dominant concentration regime, and residual diagnostics indicate unmodeled temporal structure remains. Its practical value for public-health early warning is therefore presently limited by its reduced feature scope and the resulting spike-underestimation, and its value for the multi-city spatial claim in the architecture's design has not yet been empirically tested due to the city-sort issue noted above. Establishing operational suitability will require the multivariate, per-city extension proposed in Sections 1-2, together with the broader robustness testing (regime-balanced evaluation, longer walk-forward validation, and external validation on data beyond this study window) that this baseline motivates.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Data Science Program, Faculty of Technology and Design, Bunda Mulia University, for the academic support and facilities provided throughout this research. Special appreciation is extended to the second author's academic supervisor for the guidance, constructive feedback, and continuous encouragement given during the development of the model and the writing of this paper. The authors also gratefully acknowledge Open-Meteo for providing free and openly accessible historical air quality and weather API services, which made the acquisition of the hourly PM_{2.5} and meteorological data used in this study possible. Finally, the authors thank their families, peers, and colleagues in the Data Science Program for their support and encouragement during the completion of this research.

DATA AVAILABILITY STATEMENT

The PM_{2.5}, co-pollutant, and meteorological data analyzed in this study were retrieved from the Open-Meteo Air Quality and Historical Weather APIs (<https://open-meteo.com>) for the five Jabodetabek cities and the January-July 2023 period described in Section 2.2, and can be re-queried from these public endpoints using the coordinates and date range reported there. The merged, lag-constructed dataset and model training code used to produce the results in this manuscript are not yet hosted in a public repository; they are available from the corresponding author upon reasonable request.

CODE AVAILABILITY STATEMENT

The model training and evaluation code (data preprocessing, the CNN-LSTM-Attention architecture, ensemble training, and the supplementary analyses in Sections 3.7-3.9) is available from the corresponding author upon reasonable request.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] G. Syuhada, A. Akbar, D. Hardiawan, V. Pun, A. Darmawan, S. H. A. Heryati, A. Y. M. Siregar, R. R. Kusuma, R. Driejana, V. Ingole, D. Kass, and S. Mehta, "Impacts of air pollution on health and cost of illness in Jakarta, Indonesia," *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, art. 2916, 2023, doi: 10.3390/ijerph20042916.
- [2] T. Madan, S. Sagar, Y. Singh, et al., "Hybrid deep learning model for air quality prediction and its impact on healthcare," *Sci. Rep.*, vol. 16, art. 6036, 2026, doi: 10.1038/s41598-026-36564-5.
- [3] P. Hiskiawan, Y. M. Geasela, H. Heryanto, E. Stephanie, M. Ardianti, and F. A. Sukarno, "Trustworthy data science framework for non-invasive nutritional screening using computer vision," in *Proc. 2025 Int. Conf. Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, Indonesia, 2025, pp. 1743-1748, doi: 10.1109/ICIMCIS68501.2025.11327268.
- [4] K. Zhang, X. Yang, H. Cao, J. Thé, Z. Tan, and H. Yu, "Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial-temporal attention and residual learning," *Environ. Int.*, vol. 171, art. 107691, 2023, doi: 10.1016/j.envint.2022.107691.
- [5] X. Bai, N. Zhang, X. Cao, and W. Chen, "Prediction of PM2.5 concentration based on a CNN-LSTM neural network algorithm," *PeerJ*, vol. 12, art. e17811, 2024, doi: 10.7717/peerj.17811.
- [6] D. Nelson, Y. Choi, M. Payami, D. Singh, R. Dimri, S. G. Kayastha, S. Shahriar, R. Islam, A. K. Salman, and J. Park, "Advances in air quality modeling through artificial intelligence, machine learning, and deep learning: A comprehensive review," *Sci. Total Environ.*, vol. 1021, art. 181593, 2026, doi: 10.1016/j.scitotenv.2026.181593.
- [7] M. Panja, T. Chakraborty, A. Biswas, and S. Deb, "E-STGCN: Extreme spatio-temporal graph convolutional networks for air quality forecasting," *J. R. Stat. Soc. Ser. A Stat. Soc.*, art. qnag010, 2026, doi: 10.1093/jrssa/qnag010.
- [8] N. Zaini, L. W. Ean, A. N. Ahmed, et al., "PM2.5 forecasting for an urban area based on deep learning and decomposition method," *Sci. Rep.*, vol. 12, art. 17565, 2022, doi: 10.1038/s41598-022-21769-1.
- [9] P. Hiskiawan, T. P. Wijayanti, H. Heryanto, S. Everlin, S. A. Yasodhara, and D. Alexander, "Mel-frequency cepstral coefficients and neural networks for Indonesian traditional music recognition," in *Proc. 2025 Int. Conf. Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, Indonesia, 2025, pp. 1707-1712, doi: 10.1109/ICIMCIS68501.2025.11327436.
- [10] S. Xu et al., "A novel hybrid model for six main pollutant concentrations forecasting based on improved LSTM neural networks," *Sci. Rep.*, vol. 12, art. 13957, 2022, doi: 10.1038/s41598-022-17754-3.
- [11] S. Zhou, W. Wang, L. Zhu, Q. Qiao, and Y. Kang, "Deep-learning architecture for PM2.5 concentration prediction: A review," *Environ. Sci. Ecotechnol.*, vol. 21, art. 100400, 2024, doi: 10.1016/j.ese.2024.100400.
- [12] P. Hiskiawan, C.-C. Chen, and Z.-K. Ye, "Processing of electrical resistivity tomography data using convolutional neural network in ERT-NET architectures," *Arab. J. Geosci.*, vol. 16, no. 10, art. 581, 2023, doi: 10.1007/s12517-023-11690-w.
- [13] X. Huang, M. S. Sulaiman, and M. M. Rosli, "Air quality prediction based on VMD-CNN-BiLSTM-attention," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 7, pp. 877-885, 2025, doi: 10.14569/ijacsa.2025.0160785.
- [14] D. Li, J. Liu, and Y. Zhao, "Forecasting of PM2.5 concentration in Beijing using hybrid deep learning framework based on attention mechanism," *Appl. Sci.*, vol. 12, no. 21, art. 11155, 2022, doi: 10.3390/app122111155.
- [15] J. Yuan, H. Dengxin, W. Yufeng, et al., "Attention mechanism based CNN-LSTM hybrid deep learning model for atmospheric ozone concentration prediction," *Sci. Rep.*, vol. 15, art. 21260, 2025, doi: 10.1038/s41598-025-05877-2.
- [16] X. Zou, J. Zhao, D. Zhao, B. Sun, Y. He, and S. Fuentes, "Air quality prediction based on a spatiotemporal attention mechanism," *Mobile Inf. Syst.*, vol. 2021, art. 6630944, 2021, doi: 10.1155/2021/6630944.
- [17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [18] P. Hiskiawan, M. K. Sari, R. E. Siregar, N. A. Valencia, and T. P. Wijayanti, "A deep learning data fusion approach for prostate MRI zonal segmentation using dual-channel U-Net with T2 and ADC images," in *Proc. 2026 Int. Conf. Current Research in Artificial Intelligence and Data Science (ICCRAIDS)*, Banyumas, Indonesia, 2026, pp. 1-7, doi: 10.1109/ICCRAIDS67816.2026.11519671.
- [19] B. Cui et al., "Deep learning methods for atmospheric PM2.5 prediction: A comparative study of transformer and CNN-LSTM-attention," *Atmos. Pollut. Res.*, vol. 14, no. 9, art. 101833, 2023, doi: 10.1016/j.apr.2023.101833.
- [20] C. Yin, W. Li, T. Li, et al., "Air quality prediction model based on deep learning hybrid framework," *Sci. Rep.*, vol. 16, art. 7084, 2026, doi: 10.1038/s41598-026-37896-y.