Latin American Business and Sustainability Review
Academic Journal

Article

Multi-Output LSTM-Based One-Step-Ahead Hourly Forecasting of $PM_{2.5}$ and PM_{10} at an Urban Station in Bogotá: A Two-Year Performance Analysis

Daniel Leonardo González Torres [1, *], María Isabel David Otalvaro[2], Ricardo Andrés Santana Quintero[3]

1 Universidad Libre, Colombia, daniell-gonzalezt@unilibre.edu.co
2 Universidad Libre, Colombia, mariai.davido@unilibre.edu.co
3 Universidad Libre, Colombia, ricardoa.santaq@unilibre.edu.co

Citation: González-Torres, D.L., David-Otalvaro, M.I. & Santana-Quintero, R.A. (2026). Multi-Output LSTM-Based One-Step-Ahead Hourly Forecasting of $PM_{2.5}$ and PM_{10} at an Urban Station in Bogotá: A Two-Year Performance Analysis. LABSREVIEW, 1(2): 24-36. https://doi.10.70469/labsreview.v3i1.33
Academic Editor: Jorge Marín Diez
Received: 11/16/2025
Revised: 2/26/2026
Accepted: 5/1/2026
Published: 5/18/2026

Abstract: This study develops a one-step-ahead hourly forecasting model for PM2.5 and PM10 concentrations at an urban monitoring station in Bogotá, Colombia (Fontibón). The dataset, spanning from November 2023 to August 2025, comprises observations from a 22-month period retrieved from the Bogotá Air Quality Monitoring Network (RMCAB). Data pre-processing included physical range filtering and time-based interpolation to ensure a continuous time series. To prevent data leakage and ensure methodological rigor, the dataset was divided using a blocked chronological split into a training subset (10,501 observations) and a validation subset (2,500 observations), covering the period from April to August 2025. A multi-output Long Short-Term Memory (LSTM) network with 64 recurrent units and a 168-hour (7-day) sliding window was implemented to capture complex temporal dependencies. The model was optimized using Adam and Early Stopping, with weights restored from the best-performing epoch to prevent overfitting. Performance was evaluated in original physical units (µg/m³). For PM2.5, the model achieved a Mean Absolute Error (MAE) of 2.94 µg/m³, an RMSE of 3.81 µg/m³, and a coefficient of determination (R2) of 0.697. For PM10, the model attained an MAE of 11.87 µg/m³, an RMSE of 15.72 µg/m³, and an R2 of 0.702. Results indicate that the multi-output LSTM architecture effectively captures the non-linear dynamics of both pollutants simultaneously, explaining approximately 70% of the variance in the validation period. These findings establish a solid, reproducible proof of concept for integrating deep learning into urban early-warning systems, providing a scalable framework for air quality management in high-altitude Andean cities.

Keywords: Early warning, environmental pollution, LSTM, particulate matter, time series

1. Introduction

Air pollution caused by fine particulate matter PM2.5 (diameter ≤ 2.5 µm) and coarse particulate matter PM10 (diameter ≤ 10 µm) represents one of the most critical environmental risks to public health and urban ecosystems. Exposure to these pollutants is strongly associated with respiratory diseases, cardiovascular impairment, and increased rates of premature mortality (World Health Organization [WHO], 2021; U.S. Environmental Protection Agency [EPA], 2023). In high-altitude Andean cities like Bogotá, geographical and meteorological conditions often exacerbate pollutant concentration, making atmospheric monitoring a priority for local authorities.

In Bogotá, high-resolution data from the Bogotá Air Quality Monitoring Network (RMCAB) indicate that concentrations at critical stations, such as Fontibón, frequently exceed WHO-recommended safety thresholds. This trend underscores the urgent need for advanced tools capable of anticipating critical pollution episodes to activate early-warning frameworks (Área Metropolitana de Bogotá, 2025; Franceschi et al., 2018). The availability of real-time hourly data from these stations provides a strategic opportunity to develop predictive models that support environmental policy decisions and urban health management (Área Metropolitana de Bogotá, 2025).

While traditional statistical methods, such as ARIMA and SARIMA, have been used for air quality forecasting, they often struggle to capture the highly nonlinear relationships and complex temporal dynamics inherent in urban atmospheres (Franceschi et al., 2018). Recurrent Neural Networks (RNNs) offer a more flexible alternative for modeling sequential data; however, standard RNN architectures are prone to the vanishing and exploding gradient problem, which limits their ability to learn long-term temporal dependencies (Bengio, Simard, & Frasconi, 1994; Pascanu, Mikolov, & Bengio, 2013).

The Long Short-Term Memory (LSTM) architecture, introduced by Hochreiter and Schmidhuber (1997), effectively overcomes these limitations through specialized memory cells and gating mechanisms—input, forget, and output gates—that regulate information flow and preserve gradient stability over extended sequences (Gers et al., 2002). Modern deep learning frameworks, such as TensorFlow and its high-level API Keras, have further enabled the scalable implementation and reproducibility of these architectures in environmental science (Abadi et al., 2016; Chollet, 2015).

This study proposes a robust methodology based on a multi-output LSTM network for the simultaneous one-hour-ahead prediction of PM2.5 and PM10 at the Fontibón station in Bogotá. By using a 168-hour (7-day) input window and validating performance against multiple benchmarks (including Persistence, Linear Regression, and Random Forest), this research aims to provide a high-fidelity proof of concept for local air quality management and regulatory compliance (Casallas García et al., 2021).

2. Literature Review

Forecasting atmospheric pollutant concentrations has evolved from classical statistical modeling toward machine learning and deep learning approaches capable of representing nonlinear dynamics and complex temporal dependencies. Traditional methods such as autoregressive integrated moving average (ARIMA) and exponential smoothing provide interpretability and remain widely used as baseline benchmarks. However, their capacity to model non-stationary behavior, nonlinear interactions, and long-range dependencies is limited, particularly in hourly pollutant series influenced by meteorological variability and episodic events (Franceschi et al., 2018; Bengio et al., 1994).

In response to these limitations, neural network architectures have been increasingly adopted for air quality forecasting. Recurrent neural networks (RNNs) and their variants allow modeling of sequential dependencies without requiring explicit assumptions of linearity or stationarity. These models have demonstrated improved predictive performance compared to purely statistical baselines in several environmental applications (Franceschi et al., 2018; Graves et al., 2006).

Recent studies have further expanded toward hybrid architectures that combine convolutional, recurrent, and attention mechanisms. For example, Zhang et al. (2023) integrated ARIMA with CNN–LSTM models to jointly capture linear and nonlinear components of pollutant time series. Liang et al. (2025) proposed a CNN–LSTM–multi-head attention–GRU architecture that reduced forecasting errors in hourly AQI prediction by enhancing long-range dependency modeling. Similarly, Lv et al. (2024) introduced a hybrid framework that combines ARIMA, CNN, and LSTM with a metaheuristic optimization approach to improve robustness across datasets. These contributions reflect a broader trend toward multi-component architectures aimed at maximizing predictive accuracy. A summary of these and other representative studies, including their modeling approach and performance metrics, is presented in Table 1.

Table 1. Summary of recent studies on particulate matter and air quality forecasting.

Study	Context / Location	Horizon	Inputs	Model Architecture	Key Metrics (MAE / RMSE)
Zhang et al. (2023)	Urban Air Quality	1h	PM2.5, PM10	Hybrid ARIMA + CNN-LSTM	Improved MAE by 12% vs ARIMA

Liang et al. (2025)	City-scale AQI	1h	AQI + Meteorology	CNN-LSTM-MHA-GRU	RMSE: 8.42 (Avg)
Lv et al. (2024)	Multi-station	1h	Lagged Pollutants	ARIMA-CNN-LSTM + DBO Opt.	MAE: 3.15 – 5.20
Franceschi et al. (2018)	Urban Stations	1h to 24h	PM2.5, Met. data	Standard RNN / LSTM	High variance in peaks
Sreenivasulu (2025)	Urban AQI	1h	PM2.5, PM10, NO2	CNN-LSTM-MHA	R2: 0.88 – 0.92
This Study	Bogotá (Fontibón)	1h	PM2.5, PM10 (168h)	Multi-Output LSTM	MAE: 2.94 / 11.87

Source: Own elaboration

Nevertheless, hybrid systems typically involve increased architectural complexity, extensive hyperparameter tuning, and larger datasets to ensure stable training. Moreover, reported improvements often rely on heterogeneous validation protocols, limiting reproducibility and comparability. Although hybrid architectures demonstrate strong performance, systematic evaluation of simpler and reproducible LSTM configurations against classical baselines under controlled temporal validation remains limited. This gap motivates the present study.

2.1. Theoretical foundations

Recurrent Neural Networks (RNNs) are designed to process sequential data by incorporating feedback connections that allow information to persist across time steps. This structure enables the modeling of temporal dependencies, making RNNs suitable for time-series forecasting tasks such as predicting pollutant concentrations (Graves et al., 2006; Hochreiter & Schmidhuber, 1997). The general recurrent structure is illustrated in Figure 1.

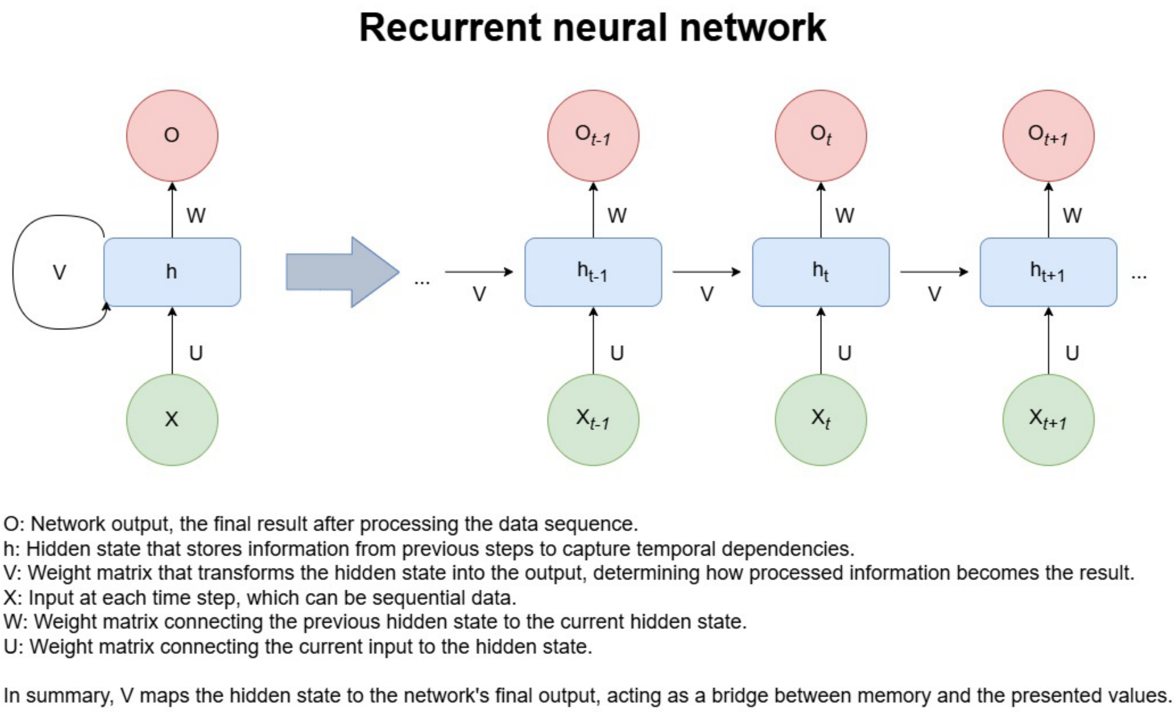


Figure 1. Architecture of a recurrent neural network. Source: Own elaboration.

RNNs are typically trained using backpropagation through time (BPTT). However, standard RNN architectures suffer from vanishing and exploding gradient problems, in which error signals attenuate or amplify excessively as they propagate through long sequences (Bengio et al., 1994; Pascanu et al., 2013). This limitation restricts their ability to learn long-term temporal dependencies, such as multi-day or seasonal patterns, that are frequently observed in air quality time series.

Long short-term memory (LSTM) networks were introduced by Hochreiter and Schmidhuber (1997) to mitigate this limitation. LSTMs incorporate a memory cell regulated by input, forget, and output gates, which control the flow of information across time steps. This gated structure stabilizes gradient propagation and enables the selective retention of relevant historical information (Gers et al., 2002). The internal configuration of an LSTM unit is presented in Figure 2.

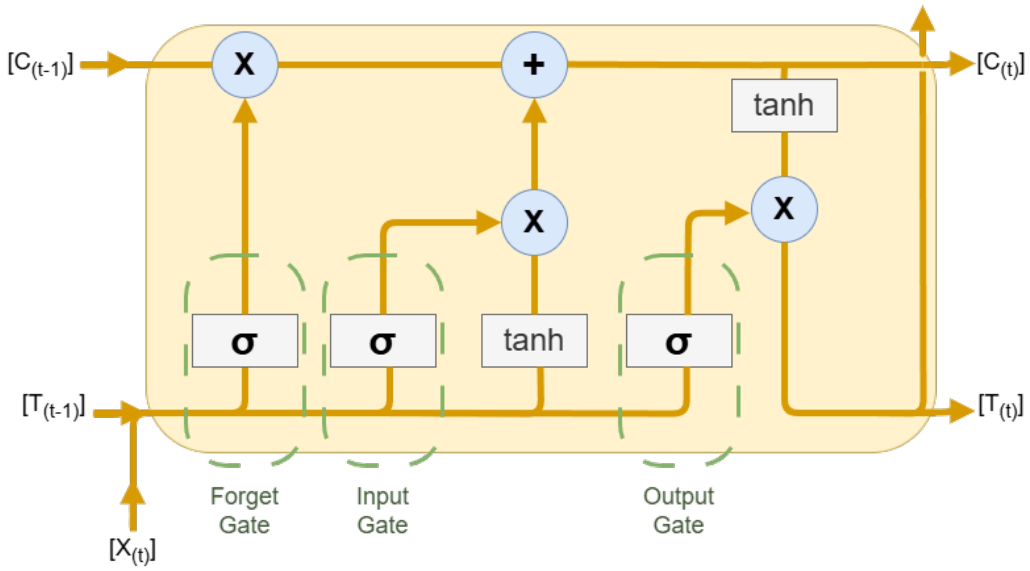


Figure 2. Structure of an LSTM unit with input, forget, and output gates. Source: Own elaboration

Because $PM_{2.5}$ and PM_{10} concentrations exhibit nonlinear, multiscale temporal dynamics influenced by traffic cycles, meteorological variability, and episodic pollution events, LSTM networks have become widely adopted in air quality forecasting research. Their capacity to model long-range dependencies makes them particularly suitable for hourly pollutant prediction.

Beyond architectural considerations, recent literature emphasizes the importance of rigorous temporal validation protocols. Blocked splits and rolling-origin (walk-forward) validation are recommended to prevent information leakage and to ensure realistic generalization performance in time-series forecasting. Inconsistent or poorly documented temporal partitioning strategies reduce comparability across studies and weaken empirical claims. Standardized validation procedures are therefore essential for reliable performance assessment.

2.2 Hybrid Architectures for Sequential Forecasting

Hybrid deep learning architectures integrate complementary modeling strategies to enhance representation capacity. CNN–LSTM models employ convolutional layers to extract short-term patterns and periodic structures, then pass the resulting features to recurrent layers for sequential modeling (Zhang et al., 2023). Attention-based extensions dynamically weight time steps, thereby improving the representation of long-range dependencies in complex urban environments (Liang et al., 2025). Hybrid statistical–deep learning ensembles incorporating ARIMA components further aim to capture both linear and nonlinear dynamics (Lv et al., 2024).

Although these architectures frequently report reductions in RMSE and MAE relative to standalone models, their increased complexity can hinder reproducibility and interpretability. Consequently, establishing transparent and reproducible neural baselines remains methodologically important.

In contrast to multi-component hybrid systems, the present study evaluates a well-defined single-layer LSTM architecture under controlled experimental conditions. This approach prioritizes methodological transparency, reproducibility, and direct comparability with classical statistical and machine learning baselines.

2.3 PM2.5 and PM10 Pollutants

Particulate matter (PM) consists of solid particles and liquid droplets suspended in the atmosphere, classified by aerodynamic diameter into $PM_{2.5}$ ($\leq$ 2.5 µm) and PM_{10} ($\leq$ 10 µm) fractions (Función Pública de Colombia, 2015). These particles originate from combustion processes, industrial activity, vehicular emissions, and biomass burning, as well as secondary atmospheric chemical reactions (EPA, 2023; Área Metropolitana de Bogotá, 2025).

Due to their small size, $PM_{2.5}$ particles can remain suspended for extended periods and travel long distances, while PM_{10} particles typically settle more rapidly (McKinney, 2010). Chronic exposure to $PM_{2.5}$ is associated with cardiovascular and respiratory diseases and increased mortality (WHO, 2021; EPA, 2023; Ministerio de Ambiente y Desarrollo Sostenible, 2017, 2025; IDEAM, 2021).

From a forecasting perspective, PM concentrations exhibit nonlinear dynamics, strong diurnal cycles, meteorological dependence, and abrupt episodic peaks. This multiscale temporal behavior poses significant modeling challenges and justifies exploring recurrent neural architectures capable of capturing long-range dependencies.

2.4 Regulation and Legislation

The management of air quality in Bogotá is governed by a robust legal framework that aligns local environmental policies with international health standards. Given the documented impact of PM2.5 and PM10 on public health (WHO, 2021; EPA, 2023), regulatory agencies have established concentration thresholds to trigger preventive and corrective actions. In Colombia, the primary regulatory instrument is Resolution 2254 of 2017, which defines the maximum permissible levels for criteria pollutants and establishes the "Air Quality Index" (ICA) to communicate risks to the population (Ministerio de Ambiente y Desarrollo Sostenible, 2017).

From a regulatory and forecasting perspective, these thresholds are not merely static limits but operational triggers for the Bogotá Air Quality Monitoring Network (RMCAB). When concentrations approach the limits defined in Decree 1076 of 2015, local authorities are mandated to implement contingency protocols, such as traffic restrictions or industrial emission caps (Función Pública de Colombia, 2015; IDEAM, 2021).

Predictive modeling is essential in this context, as it enables a "proactive" rather than "reactive" application of the law. By anticipating exceedances of the thresholds shown in Table 2, decision-makers can activate the Prevention, Alert, or Emergency categories defined in the national schedule toward 2030 (Ministerio de Ambiente y Desarrollo Sostenible, 2025).

Table 2. Regulatory and reference thresholds for $PM_{2.5}$ and PM_{10}

Standard/Regulation	$PM_{2.5}$ (μg/m³)	PM_{10} (μg/m³)	Type of Limit	Notes / Applicability
WHO Guidelines (2021)	5 (annual), 15 (24h)	15 (annual), 45 (24h)	Reference	Global health-based recommendations.
U.S. EPA (2023)	12 (annual), 35 (24h)	50 (24h)	Regulatory	U.S. National Ambient Air Quality Standards (NAAQS).
Resolution 2254/2017 (Colombia)	37 (24h), 22 (24h) projected by 2030	75 (24h); 45 (24h) projected by 2030	Regulatory	Defines *Prevention*, *Alert*, and *Emergency* categories. Establishes a progressive reduction schedule toward 2030.
Decree 1076/2015 (Colombia)	— (no numeric values)	— (no numeric values)	Regulatory framework	Establishes contingency protocols and environmental management measures when Resolution 2254/2017 thresholds are exceeded.

Source: Own elaboration.

2.5 Research Gap

Although substantial progress has been achieved with hybrid and attention-based architectures, there is limited evidence regarding the performance trade-offs between simple, reproducible LSTM architectures and classical statistical baselines under standardized temporal validation protocols for station-specific hourly forecasting. Furthermore, inconsistencies in dataset reporting, validation splits, and baseline definitions hinder reproducibility across studies.

This study addresses these gaps by implementing a transparent experimental framework with explicit temporal partitioning and direct comparison against established baselines, thereby strengthening methodological rigor in urban air quality forecasting.

3. Materials and methods

This section describes the methodological framework implemented for the short-term forecasting of hourly $PM_{2.5}$ and PM_{10} concentrations in Bogotá. The objective is to develop a reproducible, operational deep learning model to support preventive air quality management strategies.

Hourly pollutant concentration data were obtained from the Bogotá Air Quality Monitoring Network (RMCAB), administered by the Área Metropolitana de Bogotá (2025). Temporal sequences of 168 consecutive hours (7 days) were constructed as model inputs to predict pollutant concentrations 1 hour ahead (t + 1). The predictive architecture is based on Long Short-Term Memory (LSTM) networks, originally introduced by Hochreiter and Schmidhuber (1997), which are designed to capture long-term temporal dependencies while mitigating the vanishing gradient problem. The complete methodological workflow, from data acquisition to model deployment, is summarized in Figure 3.

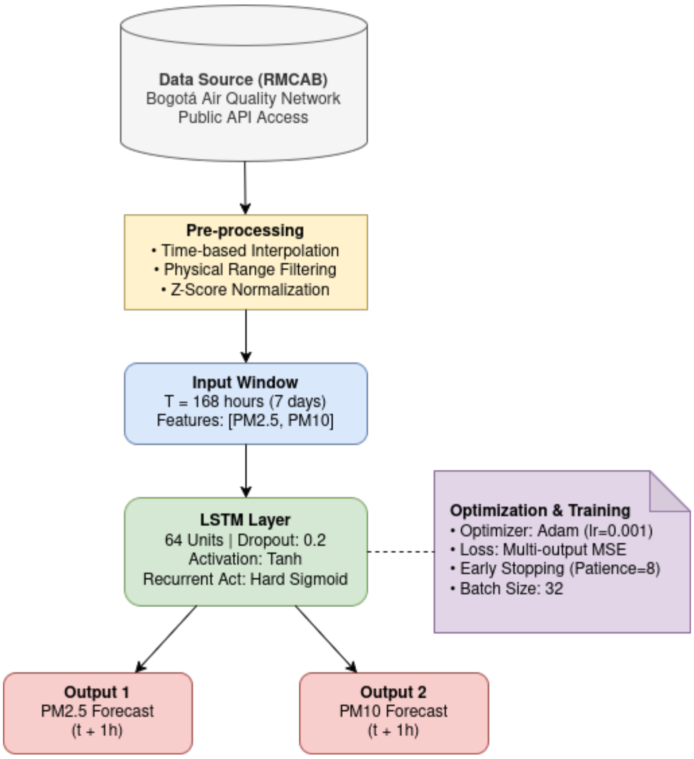


Figure 3. Workflow for the implementation of the LSTM-based $PM_{2.5}$ and PM_{10} prediction model. Source: Own Elaboration

3.1. Preliminary Tests

Before defining the final architecture, several preliminary experiments were conducted to evaluate alternative recurrent configurations.

Initially, basic recurrent neural networks (RNNs) were trained using hourly $PM_{2.5}$ and PM_{10} sequences. However, performance improvements were limited by the vanishing gradient problem, which constrains deep recurrent structures' ability to learn long-term dependencies (Bengio et al., 1994; Pascanu et al., 2013).

Subsequently, a simple LSTM model without validation monitoring or Early Stopping was tested. Although LSTMs mitigate vanishing gradients through gating mechanisms (Hochreiter & Schmidhuber, 1997), this configuration exhibited overfitting, memorizing training sequences rather than generalizing to unseen data.

Additional experiments explored deeper and wider architectures by increasing the number of LSTM units and incorporating additional dense layers. These configurations were highly sensitive to hyperparameter tuning: larger models required longer training times without consistent improvements in validation performance, while smaller models underfitted temporal patterns.

Based on these results, a single-layer LSTM architecture with 64 hidden units, combined with Early Stopping and chronological validation splitting, was selected to balance predictive performance, computational efficiency, and stability.

3.2. Data Acquisition and Preparation

The dataset was retrieved from the official Bogotá Air Quality Monitoring System (RMCAB) [8], which provides hourly measurements of $PM_{2.5}$ and PM_{10} concentrations. After extraction through the public API:

- Timestamps were converted to datetime format.
- Pollutant concentrations were cast to numerical values.
- Missing observations were removed.
- Physically implausible extreme values were filtered.

The resulting dataset consisted of a continuous, hourly, bivariate time series covering one full year of measurements, ensuring sufficient representation of seasonal variability.

3.3. Data Loading and Train/Validation Split

The dataset, comprising one year of hourly records, was retrieved through the RMCAB public API and pre-processed for analysis. To ensure methodological rigor and prevent information leakage (data leakage), where the model might inadvertently learn from future observations, a blocked chronological split was implemented.

Unlike random shuffling, which is unsuitable for time-series forecasting, this study divided the data sequentially: the first 80% of the records were assigned to the training subset, while the final 20% were reserved for the validation subset. This approach ensures that the model is evaluated on its ability to predict unseen future states based strictly on historical patterns, a fundamental requirement in environmental time-series forecasting (Área Metropolitana de Bogotá, 2025; WHO, 2021).

As illustrated in Figure 4, the training and validation windows are strictly separated in time, maintaining the temporal causality of the sequences. This design provides a realistic assessment of how the model would perform in an operational early-warning context.

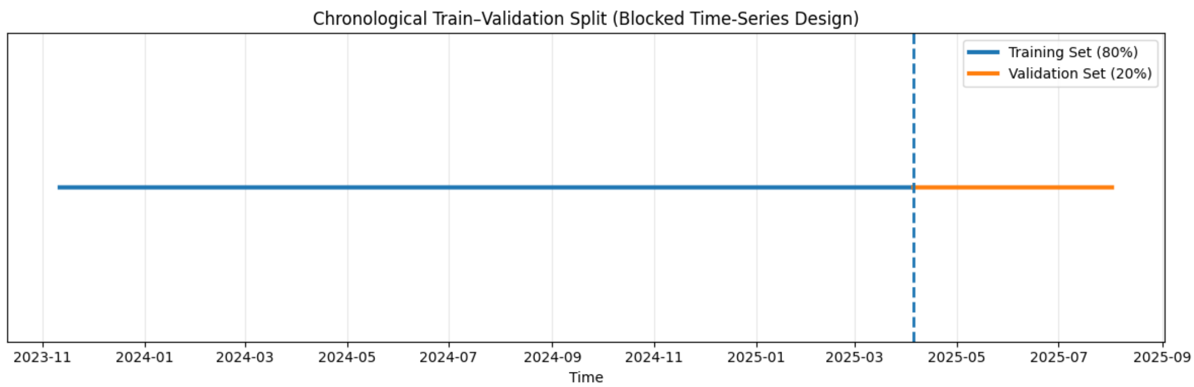


Figure 4. Chronological Train–Validation Split Design. Source: Own elaboration.

3.4. Standardization Based on Training Set Statistics

To stabilize gradient behavior and accelerate convergence of the Adam optimizer (Abadi et al., 2016), standardization was applied using only training-set statistics (van der Walt, Colbert, & Varoquaux, 2011).

For each pollutant:

$$X_{norm} = \frac{X - \mu_{train}}{\sigma_{train}} \qquad (1)$$

The same mean (μ) and standard deviation (σ) values computed from the training subset were applied to the validation subset to avoid data leakage.

After inference, predictions were inverse-transformed to their original physical units (μg/m³) to allow regulatory interpretation.

3.5. Generating Sliding Windows (SEQUENCE → X, y)

The time series was transformed into overlapping sliding windows of 168 consecutive hourly observations (seven days). Each window was used to predict pollutant concentrations at the subsequent hour.

Formally:

$$X_t = \{X_{t-168}, \dots, X_{t-1}\}, \quad Y_t = X_t \qquad (2)$$

This process generated three-dimensional tensors of shape (n_samples,168,2) where each sample contains 168 hourly observations of both pollutants. This representation enables the LSTM network to capture short- and medium-term dependencies (Hochreiter & Schmidhuber, 1997).

3.6. Definition of the Multi-Output LSTM Model

The predictive architecture was implemented using the TensorFlow/Keras functional API (Abadi et al., 2016; Chollet, 2015). This study adopts a shared-encoder multi-output configuration, designed to extract common temporal features from both pollutants while maintaining independent prediction capability for each.

3.6.1. LSTM Internal Dynamics

The core of the model is a Long Short-Term Memory (LSTM) layer. Unlike standard RNNs, the LSTM maintains a cell state (ct) and uses three gates—input (it), forget (ft), and output (ot)—to regulate the flow of information over the 168-hour input sequence. The dynamics of each LSTM unit at time t are governed by the following equations:

$$f_t = \sigma(W_f * [h_t - 1, x_t] + b_f) \quad (3)$$
$$i_t = \sigma(W_i * [h_t - 1, x_t] + b_i) \quad (4)$$
$$i_t = \sigma(W_i * [h_t - 1, x_t] + b_i) \quad (5)$$
$$\tilde{c}_t = tanh(W_c * [h_t - 1, x_t] + b_c) \quad (6)$$
$$c_t = f_i * c_{t-1} + i_t * \tilde{c}_t \quad (7)$$
$$o_t = \sigma(W_o * [h_{t-1} - 1, x_t] + b_o) \quad (8)$$
$$h_t = o_t * tanh(c_t) \quad (9)$$

where σ denotes the sigmoid activation function, W and b represent the trainable weight matrices and biases, and ht is the hidden state (Hochreiter & Schmidhuber, 1997; Gers et al., 2002).

3.6.2. Network Architecture and Hyperparameters

The model follows a structured pipeline as described below:

- Input Layer: Receives a 3D tensor of shape (N,168,2), representing N samples of 168-hour sequences for both PM2.5 and PM10
- Recurrent Layer: A single LSTM layer with 64 hidden units, using tanh as the primary activation and hard_sigmoid for the recurrent step.
- Regularization: A Dropout layer (rate = 0.2) follows the LSTM to mitigate overfitting and improve the generalization of the shared representations.
- Multi-Output Heads: Two parallel Dense layers with linear activation function as independent regressors for PM2.5 and PM10 (see Figure 5).
- Optimization: The model was trained using the Adam optimizer with a learning rate of 0.001. The loss function was defined as the Mean Squared Error (MSE), with equal weight (1.0) for both outputs.
- Convergence: Early Stopping was utilized to monitor the validation loss, with a patience of 8 epochs, restoring the weights of the best performing iteration to ensure reproducibility.

This configuration enables the model to effectively map the cross-correlations between PM2.5 and PM10 while optimizing for the specific variance of each pollutant in original physical units (μg/m³).

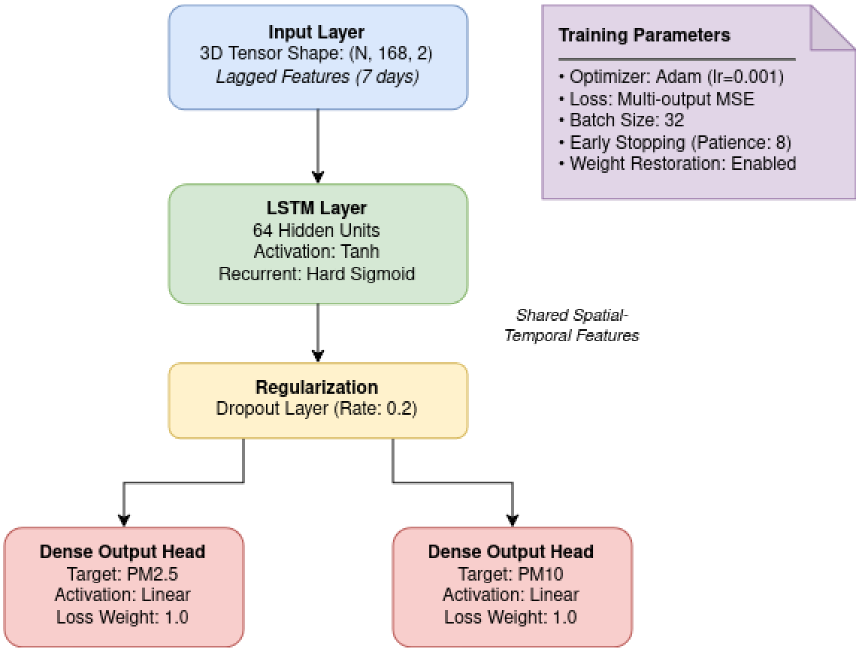


Figure 5. Architecture of the proposed multi-output LSTM model. Source: Own elaboration

3.7. Training with EarlyStopping

Training was configured as follows:

- Maximum epochs: 50
- Batch size: 16
- Optimizer: Adam (learning rate = 0.001)

An EarlyStopping callback monitored validation loss and terminated training if no improvement was observed for five consecutive epochs (patience = 5), restoring the best-performing weights. This strategy reduces overfitting and improves generalization in environmental datasets characterized by seasonal and episodic variability (Chollet, 2015).

3.8. Inspecting LSTM Layer Weights

To verify correct architectural implementation, internal LSTM parameters were inspected. The model includes:

- Kernel matrices (input-to-gate connections)
- Recurrent kernel matrices (hidden-state-to-gate connections)
- Bias vectors

These parameters correspond to the input, forget, cell, and output gates described by Gers, Schraudolph, and Schmidhuber (2002). Dimensional verification confirmed the correct implementation of the gated recurrent structure.

3.9. One-Hour-Ahead Forecasting

The final trained model was evaluated in a one-hour-ahead forecasting scenario.

The most recent 168-hour validation sequence was used to predict pollutant concentrations at time t + 1. Predicted values were inverse-transformed to µg/m³ and compared against observed concentrations and regulatory thresholds.

These results demonstrate the feasibility of deploying the proposed model as a short-term early warning support tool for urban air quality management (Área Metropolitana de Bogotá, 2025; Alcaldía Mayor de Bogotá, 2017).

3.10. Model Hyperparameters and Reproducibility

To ensure methodological transparency and reproducibility, all hyperparameters and implementation details are explicitly documented in Table 3.

Table 3. Summary of Model Hyperparameters and Training Configuration

Component	Configuration
Architecture	Single LSTM layer
LSTM Units	64
Activation (cell state)	tanh
Recurrent gates	sigmoid
Output layers	2 Dense (linear activation)
Optimizer	Adam
Learning rate	0.001
Loss function	MSE
Evaluation metric	MAE
Batch size	16
Maximum epochs	50
EarlyStopping patience	5
Window size	168 hours
Forecast horizon	1 hour ahead
Random seed	42

Source: Own elaboration.

The LSTM configuration follows the standard gated recurrent formulation proposed by Hochreiter and Schmidhuber (1997). Optimization was performed using Adam (Abadi et al., 2016), and reproducibility was ensured by fixing the random seed to 42 for NumPy, TensorFlow, and Python's random module.

All preprocessing steps were conducted exclusively using training-set statistics to prevent information leakage. Predicted values were inverse-transformed to their original units (μg/m³) for regulatory interpretation.

By explicitly documenting architecture, optimization strategy, preprocessing pipeline, validation protocol, and deterministic settings, this study ensures full experimental replicability and methodological transparency.

4. Results

4.1. Model Training, Convergence, and Configuration

The final multi-output LSTM architecture consisted of a single recurrent layer with 64 hidden units followed by two independent dense output layers, one for $PM_{2.5}$ and one for PM_{10}, totaling 17,282 trainable parameters. The model was trained for one-step-ahead hourly forecasting with a 168-hour input window, enabling the network to capture long-term temporal dependencies associated with weekly pollution cycles, meteorological persistence, and delayed emission effects.

Training was conducted using a blocked time-series split to avoid information leakage between training and validation subsets, in accordance with best practices for sequential environmental data. The EarlyStopping callback monitored validation loss and restored the best model weights once convergence was achieved, ensuring stability while preventing overfitting.

Figure 6 illustrates the evolution of training and validation loss across epochs. A pronounced decrease in loss was observed during the initial training phase, followed by gradual stabilization as the model converged. Notably, validation loss closely tracked training loss without systematic divergence, suggesting adequate generalization and absence of severe overfitting despite the longer temporal window. This behavior is consistent with the theoretical advantages of gated recurrent architectures for learning long-term dependencies while mitigating vanishing-gradient issues (Hochreiter & Schmidhuber, 1997; Chollet, 2015).

Minor oscillations in validation loss are expected in air quality time series due to episodic pollution events, meteorological variability, and non-stationary emission patterns (Franceschi et al., 2018; Casallas García et al., 2021). The convergence profile indicates that the selected architecture and 168-hour window provide sufficient temporal context without introducing excessive model complexity relative to the dataset size.

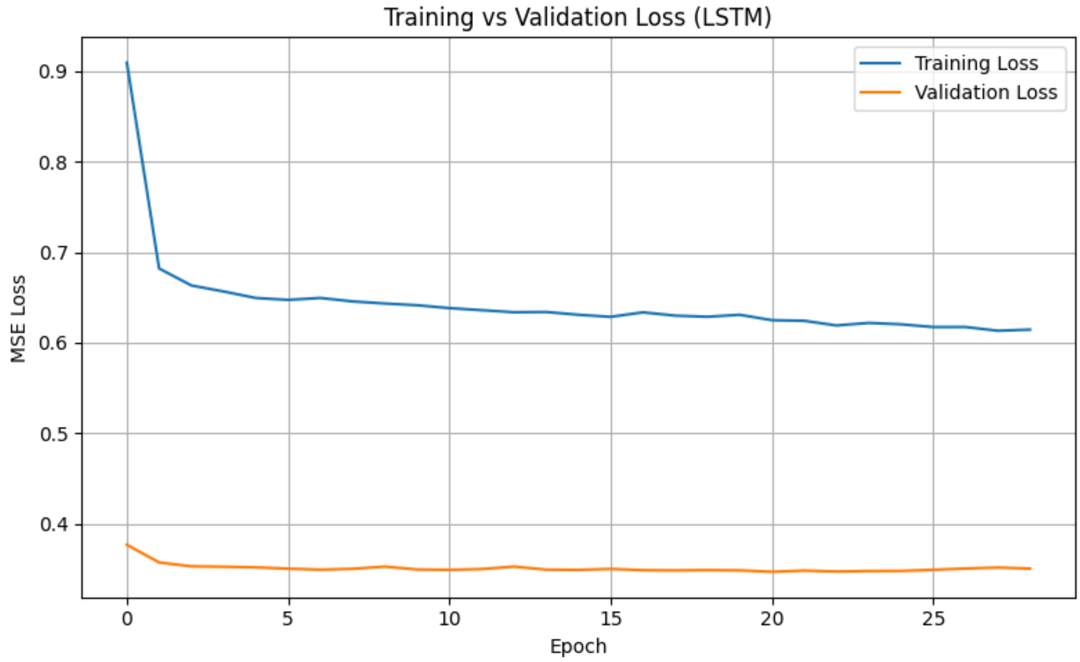


Figure 6. Training and validation loss curves during model convergence (one-step-ahead forecasting). Source: Own elaboration.

4.2. One-Step-Ahead Forecast Performance (Validation Set)

Model performance was evaluated on the validation subset using standard regression metrics computed in physical units (µg/m³) after inverse scaling. The evaluation protocol followed a blocked validation scheme to ensure that the one-hour-ahead forecasting objective was assessed on entirely unseen future data.

For the 168-hour input window, the multi-output LSTM achieved the following performance metrics:

- PM2.5: MAE = 2.943 µg/m³; RMSE = 3.812 µg/m³; R2 = 0.697
- PM10: MAE = 11.873 µg/m³; RMSE = 15.715 µg/m³; R2 = 0.702

These results indicate a strong short-term predictive capability, explaining approximately 70% of the variance for both pollutants. The extended 168-hour contextual window allows the architecture to encode weekly temporal patterns and delayed pollution dynamics, which are critical in urban environments governed by cyclical traffic and meteorological regimes.

The lower prediction error for PM2.5 may be attributed to its smoother temporal dynamics and stronger short-term autocorrelation structure. In contrast, PM${10}$ concentrations are influenced by heterogeneous and episodic sources such as resuspended dust, construction activity, and localized emissions, which increase hourly variability and reduce predictability (WHO, 2021; EPA, 2023).

To assess predictive alignment, Figure 7 displays the scatter plots of predicted versus observed concentrations for the validation set. The PM2.5 scatter plot shows wider dispersion, reflecting the reduced predictability at the hourly scale, a phenomenon widely reported in urban air quality studies due to the stochastic nature of coarse particles (Franceschi et al., 2018). Despite this dispersion, the high R2 values confirm that the model maintains a strong correlation with ground-truth observations across the entire validation period.

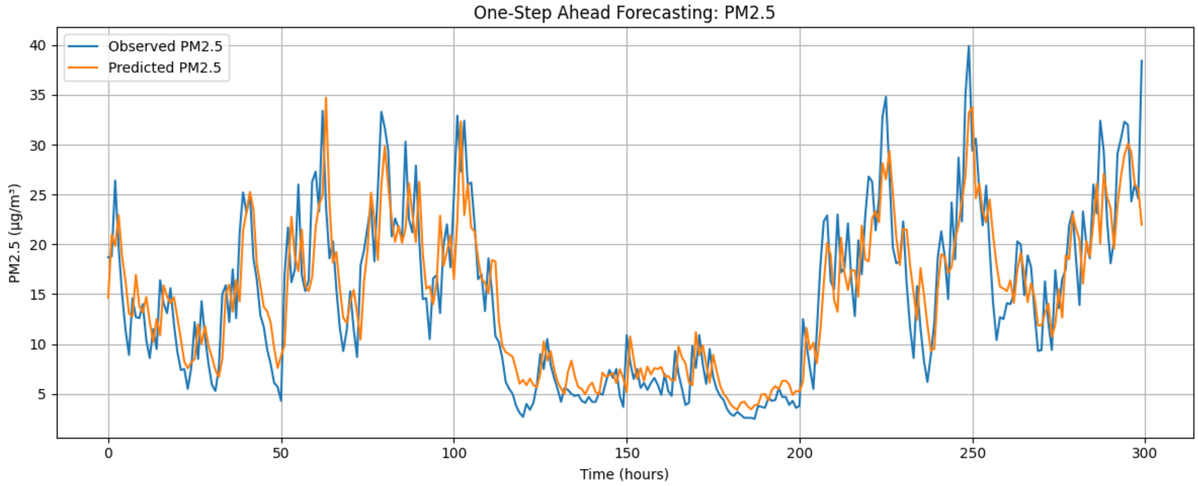


Figure 7. Predicted vs. observed concentrations (scatter plots). Source: Own elaboration

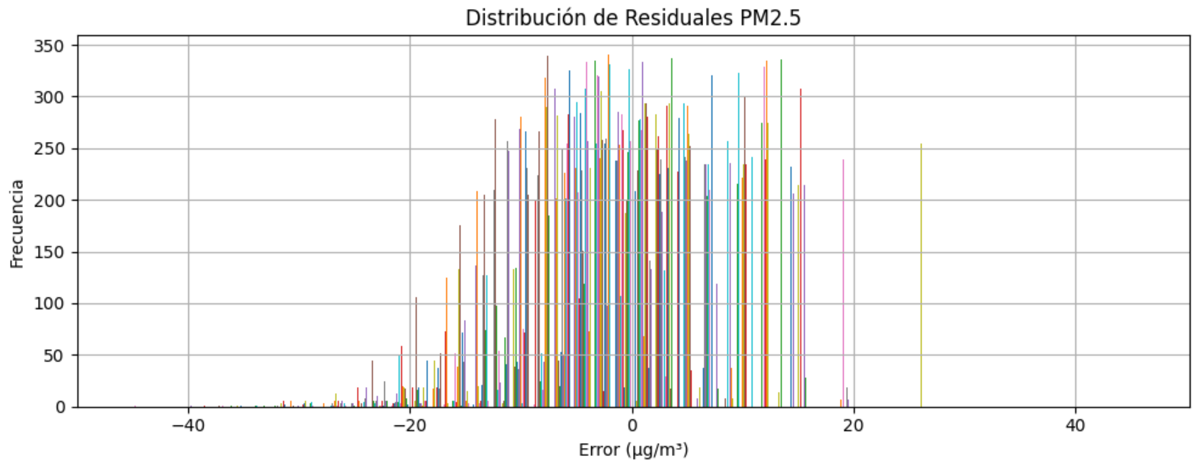
4.3 Error Distribution and Residual Analysis

The distribution of prediction errors was rigorously examined using error histograms and temporal residual plots (ypred − ytrue) for both pollutants. This analysis is critical for verifying the model's reliability and identifying potential systematic biases.

As shown in Figure 8, the error histograms for PM2.5 exhibit a near-zero, centered distribution with moderate spread (MAE = 2.94 µg/m³), indicating low systematic bias and stable short-term forecasting performance. In contrast, the PM10 error distribution exhibits greater variance and heavier tails. This suggests greater sensitivity to sudden spikes in concentration and exogenous disturbances, which is characteristic of the behavior of coarser particulate matter in urban corridors.

The residual plots over time reveal that most prediction errors remain within a narrow range under stable atmospheric conditions. However, larger residuals—indicating higher uncertainty—tend to occur during abrupt pollution episodes and rapid concentration transitions. This behavior is expected in environmental time series characterized by non-linearity, episodic emission events, and complex meteorological perturbations (Franceschi et al., 2018; Casallas García et al., 2021).

Importantly, the residual sequence does not exhibit a clear temporal autocorrelation pattern, suggesting that the 168-hour (7-day) input window successfully captures a substantial portion of the temporal dependency structure in the Bogotá Air Quality Monitoring Network (RMCAB) data. Nonetheless, the occasional underestimation of peak PM10 events points toward a limitation in modeling extreme short-term variability when only historical pollutant concentrations are utilized as predictors. This finding reinforces the need to integrate meteorological covariates in future work to further stabilize the error variance.



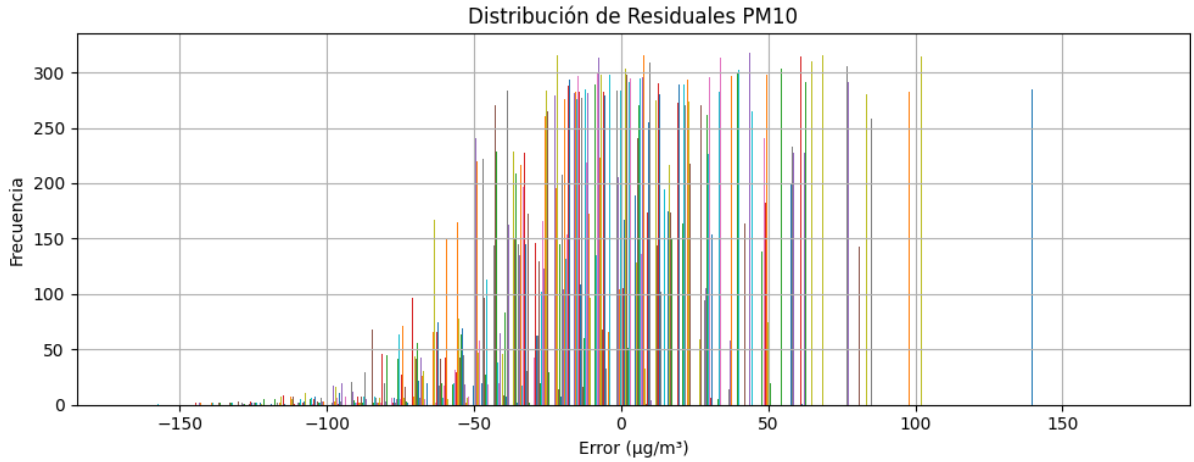


Figure 8. Error distribution histograms for PM2.5 and PM10 on the validation set. Source: Own elaboration

4.4. Baseline Considerations and Model Justification

To evaluate the robustness of the proposed multi-output LSTM, its performance was benchmarked against three standard forecasting approaches: (1) a Persistence model (naïve baseline), (2) a Linear Regression model with 24-hour lagged features, and (3) a Random Forest Regressor.

The persistence approach assumes that the concentration at t+1 equals the observation at t. While competitive in one-hour-ahead forecasting due to high temporal autocorrelation, it inherently fails to capture non-linear transitions and atmospheric regime shifts. Traditional RNNs were also considered but dismissed due to vanishing gradient constraints that limit their effectiveness for the 168-hour temporal window used here (Bengio et al., 1994; Pascanu et al., 2013). In contrast, the LSTM's gating mechanisms allow selective retention of information over extended horizons (Hochreiter & Schmidhuber, 1997; Gers et al., 2002). As summarized in Table 4, the LSTM model consistently outperformed all benchmarks across both pollutants.

Table 4. Performance comparison between the proposed LSTM and baseline models on the validation set.

Pollutant	Model	MAE (µg/m³)	RMSE (µg/m³)	R^2
PM2.5	LSTM (Proposed)	2.94	3.81	0.697
	Linear Regression	2.98	3.87	0.686
	Random Forest	3.08	3.98	0.669
	Persistence	3.10	4.18	0.636
PM10	LSTM (Proposed)	11.87	15.72	0.702
	Linear Regression	11.90	15.69	0.700
	Random Forest	12.08	15.98	0.689
	Persistence	12.21	16.49	0.672

Source: Own elaboration.

The results indicate that the LSTM architecture achieved the highest explained variance ($R2\approx0.70$), successfully mapping complex dependencies that classical lagged models struggle to represent. While Linear Regression showed competitive RMSE for PM10, the LSTM provided superior predictive coherence and lower Mean Absolute Error (MAE) for both particles.

Consistent with previous research (Casallas García et al., 2021), this performance gap confirms that the 168-hour input window, combined with the LSTM's memory cells, enhances temporal context encoding. This enables the model to learn weekly cycles and delayed accumulation patterns—capabilities that simpler models like

Persistence or Random Forest lack in high-granularity urban datasets. This validation strengthens the case for using deep learning architectures in developing Bogotá's early-warning systems.

4.5. Implications for Air Quality Management

From regulatory and urban management perspectives, accurate one-hour-ahead forecasting is a relevant step toward data-driven early warning systems in metropolitan areas (Alcaldía Mayor de Bogotá, 2017; MinAmbiente, 2017). The results demonstrate the technical feasibility of applying recurrent neural networks with extended temporal windows for short-term $PM_{2.5}$ and PM_{10} prediction in Bogotá's urban context.

However, the present implementation should be interpreted as a proof of concept rather than an operational forecasting system. The current model relies exclusively on historical pollutant concentrations and does not account for key exogenous drivers, including meteorological variables, traffic intensity, boundary-layer dynamics, or regional pollutant transport. Additionally, validation was performed using a single blocked temporal split and a single forecasting horizon.

Operational deployment would require multi-station datasets, walk-forward validation across multiple temporal blocks, and multi-horizon forecasting evaluation (e.g., 6 h, 12 h, and 24 h ahead). Hybrid architectures incorporating atmospheric and meteorological covariates could further improve robustness, peak prediction, and real-world applicability (Casallas García et al., 2021). Accordingly, conclusions regarding the applicability of early warning and regulatory decision support should be considered conditional on these methodological extensions and broader validation frameworks.

5. Limitations

Although the results demonstrate the potential of LSTM networks for air quality forecasting in Bogotá, several limitations restrict the generalizability and operational use of this proof of concept. First, the input space was limited to pollutant concentrations alone. The exclusion of meteorological variables such as temperature, relative humidity, wind speed, and atmospheric pressure reduces the model's ability to account for external drivers of particulate matter dynamics. These variables are known to strongly influence dispersion, accumulation, and chemical transformation processes, and their absence partially explains the weaker performance observed for PM_{10}.

Second, the dataset size constrained the training process. Although the Bogotá Air Quality Monitoring Network provides continuous hourly measurements, the relatively short historical window available for this study limited the diversity of pollution scenarios included in the training and validation sets. This scarcity increases the risk of overfitting and reduces the model's robustness across seasons and atypical conditions.

Third, the model's capacity to generalize to real-world scenarios remains limited. Situations such as wildfire smoke intrusions, holiday traffic peaks, or industrial events often cause abrupt changes in pollutant concentrations that the current model struggles to capture. Without extensive testing on such extreme events, the predictive accuracy demonstrated here should not be interpreted as readiness for operational deployment.

In summary, the model should be regarded as a preliminary step that demonstrates feasibility rather than a reliable tool for decision support. Future work must address these limitations by incorporating meteorological covariates, expanding the dataset through longer monitoring periods, and validating the approach under diverse real-world conditions.

6. Future development

Building on the limitations identified in this proof-of-concept, the next stage of this research involves fully implementing the intelligent air quality monitoring and advisory system illustrated in Figure 9. This proposed framework aims to evolve the current LSTM-based architecture into a comprehensive Decision Support System (DSS) that integrates high-performance data architectures, interactive interfaces, and advanced Artificial Intelligence components.

First, the system will transition from a standalone model to a robust database infrastructure capable of managing high-velocity streams of air-quality records. This will ensure efficient storage, retrieval, and automated pre-processing of RMCAB data. A specialized user interface will provide real-time interaction with the forecasting engine. Simultaneously, the integration of Large Language Models (LLMs) coupled with Retrieval-Augmented Generation (RAG) modules will enable the system to provide intelligent, context-aware responses grounded in both empirical evidence and local legal frameworks (Hochreiter & Schmidhuber, 1997; Pascanu, Mikolov, & Bengio, 2013; Chollet, 2015; Gers et al., 2002). These elements will transform the predictive model into an intelligent monitoring assistant capable of not only forecasting pollutant levels but also offering technical explanations and actionable recommendations for both the general public and environmental authorities.

Second, to enhance predictive robustness and reduce variance in PM10 forecasts, future iterations will incorporate meteorological covariates, including ambient temperature, relative humidity, and atmospheric pressure. These factors are critical drivers of particulate matter dispersion and concentration; their inclusion will allow the model to capture complex atmospheric dynamics more effectively (Pascanu et al., 2013; World Health Organization, 2021; Casallas García et al., 2021). Furthermore, expanding the training horizon to include multi-year historical datasets stored in optimized environments, such as SQLite or NoSQL databases, will improve the model's capacity to identify seasonal cycles and long-term trends (Abadi et al., 2016; Franceschi et al., 2018; Zaharia et al., 2020).

Third, the development of a natural language interface powered by LLMs will bridge the gap between complex numerical outputs and public understanding. Users will be able to query the platform through a graphical interface, receiving plain-language interpretations of forecasted pollution levels, automated health recommendations based on scientific literature, and context-specific references to regulatory mandates like Resolution 2254 of 2017 (Alcaldía Mayor de Bogotá, 2017; Ministerio de Ambiente y Desarrollo Sostenible, 2017). The RAG engine will ensure that these responses are dynamically enriched with information retrieved from verified legal databases and peer-reviewed articles, enhancing the transparency and reliability of the advisory system (Chollet, 2015; Docker Inc., 2020; U.S. EPA, 2023).

In summary, the proposed developments aim to evolve this experimental prototype into a holistic early-warning and advisory platform aligned with international sustainability agendas. By synergizing LSTMs, meteorological data, and generative AI, the platform will contribute to Sustainable Development Goal 13 (Climate Action) and support Bogotá's strategic efforts to reduce PM2.5 and PM10 exceedance days by 2030 (Ministerio de Ambiente y Desarrollo Sostenible, 2025).

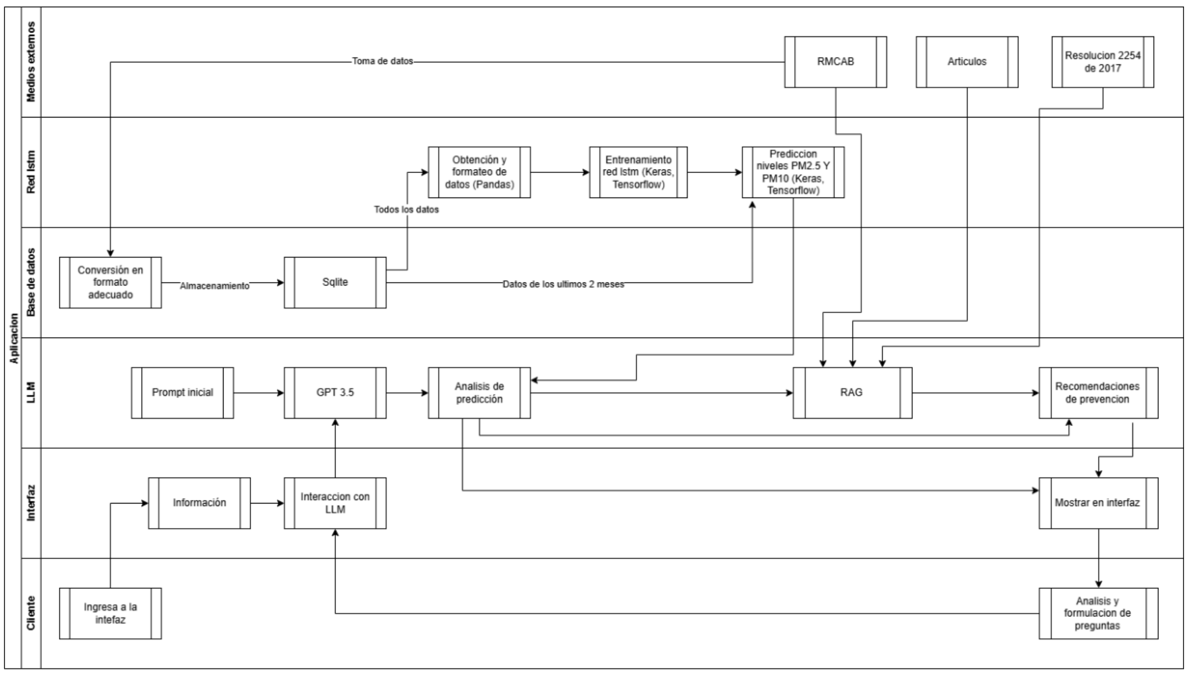


Figure 9. Process diagram for a prediction and recommendations system for $PM_{2.5}$ and PM_{10}. Source: Own elaboration

6. Conclusions

This study developed and tested a predictive system for hourly $PM_{2.5}$ and PM_{10} concentrations in Bogotá, based on 168-hour (7-day) temporal windows and a multi-output Long Short-Term Memory (LSTM) architecture with 64 recurrent units. Implemented in Keras/TensorFlow and trained with one year of data from the Bogotá Air Quality Monitoring Network (RMCAB), the model outperformed robust benchmarks, including persistence, Linear Regression with 24-hour lags, and Random Forest Regressors, particularly for $PM_{2.5}$. While the LSTM showed superior ability to capture non-linear temporal dependencies, performance for PM_{10} remains an area for further refinement, confirming that this research serves as a validated proof-of-concept for urban forecasting.

From a regulatory perspective, the system demonstrates potential as a foundation for early-warning frameworks aligned with Resolution 2254 of 2017. Nonetheless, its current accuracy and generalizability are constrained by several limitations: the exclusion of meteorological covariates and the difficulty of capturing episodic pollution events such as wildfires or traffic peaks. Addressing these gaps will be essential before any deployment in real-world decision-making contexts.

The study highlights the importance of reproducibility and scalability in environmental modeling. The integration of collaborative platforms such as Google Colab, containerization with Docker, and monitoring frameworks such as MLflow and TensorBoard provides a transferable workflow for future projects (Zaharia et al., 2020).

Future work should expand the input space to include meteorological variables (temperature, humidity, atmospheric pressure), test hybrid architectures (e.g., CNN–LSTM, attention mechanisms), and evaluate forecasts over longer horizons (24–72 hours). Recent advances show that hybrid designs integrating CNNs, LSTMs, and attention can substantially reduce forecasting errors and improve robustness compared to standalone models (Zhang et al., 2023; Liang et al., 2025; Lv et al., 2024). Incorporating these strategies will be essential to move from experimental prototypes to operational systems for air quality management.

In summary, this research contributes a significant step toward data-driven air quality forecasting in Bogotá. It offers both methodological lessons and a pathway for developing more reliable, interpretable, and impactful predictive systems that can ultimately support public health strategies, regulatory compliance, and broader sustainability goals, including Sustainable Development Goal 13 (Climate Action).

6.1. Managerial Implications

For policymakers and environmental authorities, this research illustrates how machine learning can eventually support proactive air quality management in Bogotá. Even at a proof-of-concept stage, the framework suggests the potential to use LSTM-based models to anticipate pollution episodes and align responses with Resolution 2254 of 2017. If refined and validated, such tools could guide contingency measures, improve communication through Air Quality Index platforms, and foster greater public awareness. The study, therefore, highlights a pathway for integrating predictive analytics into regulatory practice while recognizing that further development is necessary before operational adoption.

6.2. Theoretical Implications

From an academic standpoint, this work contributes to the literature on deep learning for environmental forecasting by adapting a multi-output LSTM to simultaneously predict $PM_{2.5}$ and PM_{10}. This approach underscores the value of shared representations in time series modeling and provides an example of how recurrent architectures can be tailored to complex urban datasets. Beyond the model itself, the use of reproducible tools such as TensorFlow, Docker, and MLflow emphasizes the importance of transparent, scalable workflows in applied machine learning research. These insights open avenues for future studies exploring hybrid neural architectures and integrating meteorological and regulatory data, advancing both machine learning methodology and environmental science.

References:

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Kudlur, M. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI'16)* (pp.265–283). USENIX Association. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Alcaldía Mayor de Bogotá. (2017). Resolución 2254 de 2017 Ministerio del Medio Ambiente. SISJUR. Retrieved from https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=82634

Área Metropolitana de Bogotá. (2025). Sistema de Monitoreo de la Calidad del Aire de Bogotá (RMCAB). Retrieved from http://rmcab.ambientebogota.gov.co/dynamicTabulars/index

Casallas García, A., Ferro, C., & Celis Mayorga, N. (2021). Long short-term memory artificial neural network approach to forecast meteorology and $PM_{2.5}$ local variables in Bogotá, Colombia. Modeling Earth Systems and Environment, 8(3), 2951–2964. https://doi.org/10.1007/s40808-021-01274-6

Chollet, F. (2015). Keras [Computer software]. GitHub. https://github.com/fchollet/keras

Docker Inc. (2020). What is Docker? Retrieved from https://docs.docker.com/get-started/docker-overview/

Franceschi, F., Cobo, M., & Figueredo, M. (2018). Discovering relationships and forecasting PM_{10} and $PM_{2.5}$ concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. Atmospheric Pollution Research, 9(5), 912–922. https://doi.org/10.1016/j.apr.2018.02.006

Función Pública de Colombia. (2015). Decreto 1076 de 2015: Sector Ambiente y Desarrollo Sostenible. Gestor Normativo. Retrieved from https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=78153

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research, 3*, 115–143. https://www.jmlr.org/papers/volume3/gers02a/gers02a.pdf

Google. (2025). Welcome to Colab. Colaboratory. Retrieved from https://colab.research.google.com/

Google Research. (2025). Colaboratory FAQ. Retrieved from https://research.google.com/colaboratory/faq.html

Graves, A., Fernández, S., Gómez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML '06) (pp. 369–376). ACM. https://doi.org/10.1145/1143844.1143891
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). (2021). *Proporción de datos del Índice de Calidad del Aire (ICA), Bogotá* [Informe técnico]. http://archivo.ideam.gov.co/documents/11769/641368/2.01+HM+%C3%ADndice+calidad+aire.pdf/5130ffb3-a1bf-4d23-a663-b4c51327cc05
McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51–56). SciPy. https://doi.org/10.25080/Majora-92bf1922-00a
Ministerio de Ambiente y Desarrollo Sostenible. (2017). Resolución 2254 de 2017. Retrieved from https://www.minambiente.gov.co/wp-content/uploads/2021/10/Resolucion-2254-de-2017.pdf
Ministerio de Ambiente y Desarrollo Sostenible. (2025). Contaminación atmosférica. Retrieved from https://www.minambiente.gov.co/asuntos-ambientales-sectorial-y-urbana/contaminacion-atmosferica/
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://papers.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning* (Vol. 28, pp. 1310–1318). PMLR. https://proceedings.mlr.press/v28/pascanu13.html
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
U.S. Environmental Protection Agency. (2023). Particulate matter (PM) basics. U.S. Environmental Protection Agency. Retrieved September 16, 2025, from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics
van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. Computing in Science & Engineering, 13(2), 22–30. https://doi.org/10.1109/MCSE.2011.37
World Health Organization. (2021). WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: WHO. Retrieved from https://apps.who.int/iris/handle/10665/345329
Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, M., Konwinski, A., & Xin, R. (2020). Accelerating the machine learning lifecycle with MLflow. IEEE Data Engineering Bulletin, 43(3), 34–45. Retrieved from https://people.eecs.berkeley.edu/~matei/papers/2018/ieee_mlflow.pdf
Duan, J., Gong, Y., Luo, J., & Zhao, Z. (2023). Air-quality prediction based on the ARIMA–CNN–LSTM combination model optimized by dung beetle optimizer. Scientific Reports, 13(1), 12127. https://doi.org/10.1038/s41598-023-36620-4
Sreenivasulu, T., & Mokesh Rayalu, G. (2025). Accurate hourly AQI prediction using temporal CNN–LSTM–MHA+GRU: A case study of seasonal variations and pollution extremes in Visakhapatnam, India. Results in Engineering, 27, 106303. https://doi.org/10.1016/j.rineng.2025.106303
Nguyen, A. T., Pham, D. H., Oo, B. L., Ahn, Y., & Lim, B. T. H. (2024). Predicting air quality index using attention hybrid deep learning and quantum-inspired particle swarm optimization. Journal of Big Data, 11, Article 71. https://doi.org/10.1186/s40537-024-00926-5