ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.2736

PM2.5 Forecasting in an Urban Environment: A Multi-Horizon Comparison of Statistical and Deep Learning Models in Santiago, Chile

Eliecer Peña-Ancavil[1]; Carla Angulo[1]; Patricio Puebla[1]; Cristopher Contreras[1]; Valentina Da-Costa-Vellozo[1]
[1]Universidad Bernardo O'Higgins, Chile. *eliecer.r.pena.ancavil@gmail.com*, *carla.angulo@ubo.cl*, *patricio.puebla@ubo.cl*, *cristopher.contreras.informatica@gmail.com*, *dvcvale@gmail.com*

Abstract—Forecasting fine particulate matter (PM2.5) concentrations is an important task in urban environmental management because of its effects on human health and its association with critical air pollution episodes. This study presents a controlled univariate multi-horizon comparison between classical statistical models and deep learning for daily PM2.5 forecasting at an urban monitoring station in Santiago, Chile. The analysis used an official time series from the Sistema de Información Nacional de Calidad del Aire (SINCA) covering 2010–2025. Four approaches were evaluated: a weekly Seasonal Naïve baseline, ARIMA, SARIMA, and a Long Short-Term Memory (LSTM) network. Forecasts were assessed at 1, 7, 14, 21, and 28 days ahead using a chronological train-validation-test split and temporally consistent model selection and evaluation procedures. Performance was measured with RMSE and R^2. The results showed that LSTM achieved the best overall performance across all forecasting horizons, with RMSE values ranging from 7.61 to 12.98, while the Seasonal Naïve baseline showed the weakest results. LSTM also obtained the highest explanatory performance, consistently outperforming ARIMA, SARIMA, and the baseline. These findings indicate that even under a strictly univariate formulation, model choice has a substantial effect on PM2.5 forecasting accuracy.

Keywords—PM2.5 forecasting, air quality, time series forecasting, deep learning, LSTM.

I. Introduction

Fine particulate matter PM2.5 consists of particles with an aerodynamic diameter less than or equal to 2.5 μm, capable of penetrating deeply into the respiratory system and associated with adverse effects on human health [1]. Due to their size and atmospheric persistence, these particles can reach high concentrations in densely populated urban environments, particularly in cities with intense vehicular, industrial, and residential activity [2]. In Chile, the management of critical pollution episodes also considers operational PM2.5 ranges of alert between 80 and 109 $\mu g/m^3$, pre-emergency between 110 and 169 $\mu g/m^3$, and emergency for concentrations greater than or equal to 170 $\mu g/m^3$, which reflects the environmental and public health relevance of their monitoring and anticipation [3].

In Santiago, Chile, air pollution has historically constituted a major environmental problem [4] [5], especially during the colder months, when atmospheric ventilation decreases and pollutant accumulation is favored. In this context, the ability to anticipate the evolution of PM2.5 is relevant for air quality surveillance, the implementation of preventive measures, and the analysis of urban pollution trends. The availability of official records from the National Air Quality Information System of Chile (Sistema de Informacion Nacional de Calidad del Aire, SINCA) makes it possible to address this problem through temporal modeling tools applied to long and environmentally representative historical series [6].

Among the available strategies for environmental time series forecasting, classical statistical models continue to be widely used because of their interpretability and theoretical foundation, whereas deep learning approaches have gained interest because of their ability to model nonlinear relationships and more complex temporal dependencies [7] [8]. However, the relative performance of these model families depends both on the nature of the series and on the forecasting horizon considered. In air quality problems, this distinction is particularly important, since a model may perform adequately at short horizons and lose predictive capacity when longer anticipation is required [1] [4] [9].

This study provides a controlled, univariate, and multi-horizon methodological comparison of daily PM2.5 forecasting methods using official data from an urban station in Santiago, Chile, for the period 2010–2025. Specifically, a weekly seasonal naïve baseline, an AutoRegressive Integrated Moving Average (ARIMA) model, a Seasonal AutoRegressive Integrated Moving Average (SARIMA) model, and an LSTM network are compared at forecasting horizons of 1, 7, 14, 21, and 28 days. The study seeks to analyze how the relative performance of these approaches changes as the forecasting horizon increases, while maintaining a homogeneous evaluation framework over an official environmental monitoring series. The results show that LSTM achieved the best overall performance, with Root Mean Square Error (RMSE) values between 7.61 and 12.98 and R^2 values between 0.754 and 0.284, consistently outperforming the statistical models and the baseline. The contribution of this work does not lie in proposing a new forecasting architecture, but in providing a controlled and chronologically consistent benchmark for comparing established statistical and deep learning models under homogeneous conditions. By using official long-term data from Santiago, Chile, and evaluating multiple forecasting horizons, the study offers a reproducible reference for assessing how model choice affects daily PM2.5 prediction in an urban Latin American context.

II. Methodology

The present study was designed as a methodological comparison between classical statistical approaches and deep learning for the univariate forecasting of daily PM2.5 concentrations in an urban context. From the perspective of environmental engineering, the interest lies not only in the abstract predictive performance of the models, but also in their ability to represent the temporal evolution of an atmospheric pollutant of high public health and environmental relevance across different forecasting horizons. For this purpose, a chronologically consistent experimental protocol was defined, based on a single official historical PM2.5 series, so that all methods were evaluated under the same temporal conditions and on the same target variable.

Since the objective of the study was to compare forecasting models on a daily PM2.5 series, the target variable was constructed as the calendar-day average of the hourly concentrations recorded in each temporal subset. This decision simplifies the modeling problem and maintains consistency with the univariate approach adopted, while also allowing a homogeneous comparison among methods with different assumptions and internal mechanisms. However, it should be noted that the regulatory thresholds associated with critical PM2.5 pollution episodes in Chile are defined on the basis of 24-hour moving concentrations, that is, averages calculated over consecutive 24-hour windows rather than strictly over calendar-day averages [3]. Therefore, in this work these thresholds are considered as a contextual reference for interpreting concentration magnitudes, but not as an exact reconstruction of the official operational criterion used to determine critical episodes.

A. Data Source

The data used in this study were obtained from the SINCA, the official environmental monitoring platform in Chile [6]. In particular, hourly PM2.5 concentration records from the La Florida (Acreditada) station were used. This station is located in the commune of La Florida, in the Santiago Metropolitan Region, and is operated by the Ministry of the Environment (Ministerio del Medio Ambiente) [10]. According to the official station record, it has been accredited since September 27, 2024, receives data online, and has a reported start of operation on March 24, 1997. For the MP2.5 parameter, the reported measurement technique is beta radiation attenuation, with a range from 4 to 1000 $\mu g/m^3$. The station record also reports historical availability of hourly MP2.5 records since January 1, 2000.

The period analyzed in this study ranged from 01-01-2010 to 31-12-2025, covering the years 2010 to 2025 and providing a sufficiently long series for model fitting, selection, and final evaluation. The choice of SINCA is justified by the fact that it provides official, continuous, and standardized air quality records, making it an appropriate source for temporal modeling studies in environmental engineering. Although the original database included other atmospheric pollution variables, the present study was deliberately restricted to the historical PM2.5 series as the sole input and output variable. This methodological decision was intended to ensure comparability among models, preventing differences in the use of exogenous variables from introducing structural advantages for any particular family of methods. Consequently, the problem was formulated as a strictly univariate forecasting task based on a real daily environmental monitoring series.

As a contextual reference for the environmental interpretation of the series, the operational ranges commonly associated with critical PM2.5 episodes in Chile may be mentioned: alert between 80 and 109 $\mu g/m^3$, pre-emergency between 110 and 169 $\mu g/m^3$, and emergency for values greater than or equal to 170 $\mu g/m^3$. In this work, these levels [3] were not used as a training variable or as a classification criterion, but only as an interpretive framework for the magnitude of certain observed concentrations.

B. Construction of the Daily Series and Preprocessing

The original dataset was loaded from separate files for training, validation, and test, retaining only the common columns available across the three subsets. Subsequently, it was verified that the temporal index of each subset corresponded to a DatetimeIndex, records with invalid dates were removed, the observations were sorted chronologically, and the variables were converted to numeric format. These operations ensured temporal consistency and reduced potential errors derived from heterogeneous formats in later stages of the pipeline.

Let x_t denote the hourly PM2.5 concentration observed at hour t, and let D_d denote the set of hours belonging to calendar day d. In addition, let n_d be the number of hourly observations available on that day. From the hourly PM2.5 series, a daily series was constructed by taking the average over each calendar day, defined as:

$$y_d = \frac{1}{n_d} \sum_{t \in D_d} x_t \tag{1}$$

Where y_d represents the average daily PM2.5 concentration for day d. This daily series constituted the common basis used in the final comparison among models.

In addition to the original daily version, a regularized daily version was generated through reindexing to daily frequency and missing value treatment within each temporal block. For this purpose, time-based interpolation was applied, followed by forward filling and backward filling, yielding a continuous series suitable for the construction of sliding windows in the LSTM model. In operational terms, this stage made it possible to obtain a regular representation of the signal without altering the chronological logic of the main partition. The complete series was retained both in its original version and in its filled version in order to preserve methodological traceability between the aggregated signal and the signal ultimately used in the experiments.

Fig. 1 shows the daily behavior of PM2.5 in the training, validation, and test subsets. This visualization shows the marked interannual variability of the series, the presence of seasonal cycles, and the temporal continuity of the experimental partition. From an environmental perspective, the figure also shows that the highest concentrations tend to cluster in certain periods of the year, which reinforces the relevance of addressing the problem through temporal modeling techniques.

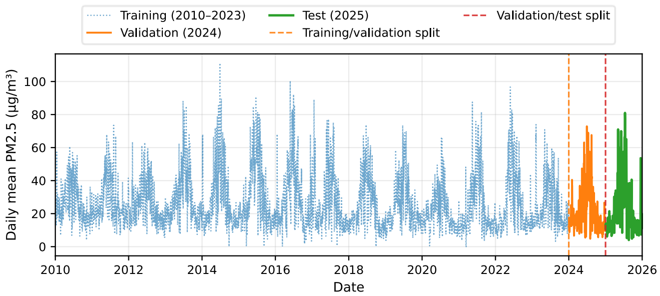

Fig. 1 Daily behavior of PM2.5 in the training, validation, and test subsets, constructed from the calendar-day average of hourly concentrations.

C. Temporal Partition of the Experiment

To preserve the temporal structure of the problem, the series was not randomly divided. Instead, a chronological partition into three subsets was employed: training, validation, and test. The training set covered the period 2010 to 2023, the validation set corresponded to the year 2024, and the test set corresponded to the year 2025. This strategy separated the stages of model fitting, configuration selection, and final out-of-sample evaluation.

The choice of a chronological partition responds to the prospective nature of the environmental forecasting problem. In real air quality applications, predictions must be issued using only information available up to the current time, so a random partition would introduce temporal leakage and prevent a realistic evaluation [9] [11]. Under this design, the validation set was used to select configurations or compare internal variants of each model family, whereas the test set was reserved for the final evaluation.

The forecasting horizons considered were 1, 7, 14, 21, and 28 days. If h denotes the forecasting horizon, then the set of evaluated horizons was $\mathcal{H} = \{1, 7, 14, 21, 28\}$. This selection made it possible to study both short-term prediction and the progressive degradation of performance as the required anticipation increased. From an environmental perspective, these horizons cover scenarios ranging from rapid response to time windows of interest for the preventive monitoring of pollution trends.

D. Univariate Formulation and Comparability Criterion

The study was formulated in a strictly univariate manner. This means that all models used only the past history of the daily PM2.5 series to generate future predictions, without incorporating meteorological variables, other stations, or additional pollutants as model inputs. The main reason for this decision was to ensure a homogeneous basis for comparison among methods of different nature.

Let y_t denote the daily PM2.5 concentration at time t, and let h be a forecasting horizon belonging to $\mathcal{H}$. Under a general univariate formulation, the forecast can be represented as a function of the past values of the series itself:

$$\hat{y}_{t+h} = f(y_t, y_{t-1}, y_{t-2}, \dots, y_{t-L+1}) \quad (2)$$

Where $\hat{y}_{t+h}$ is the predicted value for horizon h, $f(.)$ represents the prediction rule of the model, and L corresponds to the effective history length used.

In particular, the inclusion of exogenous variables could have benefited certain approaches unevenly, especially more flexible models or those with greater representational capacity, making it difficult to attribute the observed differences to the temporal modeling mechanism itself [4] [9]. In contrast, the univariate approach forces all methods to extract information exclusively from the historical dynamics of the pollutant, which makes the comparison methodologically cleaner [12] [13]. Under this formulation, performance differences mainly reflect the ability of each approach to capture temporal dependence, seasonality, smoothing, and nonlinearity present in the PM2.5 series [1] [9]. Thus, this section frames the study as a controlled comparison of model families for daily PM2.5 forecasting using an official urban Chilean series.

E. Evaluated Models

A weekly Seasonal Naive model was used as the baseline reference. This approach assumes that the most recent weekly pattern will repeat itself and, since the series was analyzed at daily frequency, a seasonal period of $s = 7$ days was fixed [12]. In its simplest form, the forecast was defined as:

$$\hat{y}_{t+h} = y_{t+h-s} \quad (3)$$

Where $\hat{y}_{t+h}$ is the forecasted value for horizon h, y_{t+h-s} is the observation corresponding to the same relative point within the previous weekly cycle, and s is the fixed seasonal period. Its main value in this study was to act as an interpretable and environmentally reasonable baseline for a daily series with a possible intra-week structure [14].

The ARIMA model was included as a representative of classical statistical approaches for time series [2] [9] [14] [15]. Let B be the lag operator, such that $By_t = y_{t-1}$; let $\phi(B)$ be the autoregressive polynomial, $\theta(B)$ the moving average polynomial, d the order of differencing, and ε_t the error term. In general notation, the $ARIMA(p, d, q)$ model can be written as:

$$\phi(B)(1 - B)^d y_t = c + \theta(B)\varepsilon_t \quad (4)$$

Where c is a constant. In this study, ARIMA was used to capture the non-seasonal temporal structure present in the daily PM2.5 series.

The SARIMA model extends the ARIMA approach by incorporating explicit seasonal components [7] [14] [15]. Let $\Phi(B^s)$ and $\Theta(B^s)$ denote the seasonal autoregressive and seasonal moving average polynomials, D the seasonal differencing order, and s the seasonal period. Its general formulation can be expressed as:

$$\Phi(B^s)\phi(B)(1 - B)^d(1 - B^s)^D y_t = c + \Theta(B^s)\theta(B)\varepsilon_t \quad (5)$$

This family is particularly relevant when the series exhibits persistent periodic patterns, as often occurs in environmental signals affected by weekly or seasonal cycles.

Finally, a Long Short-Term Memory (LSTM) network was evaluated, which is a recurrent architecture designed to model temporal dependencies and nonlinear relationships in sequences [1] [7] [15]. In this case, the univariate series was transformed into a supervised dataset using sliding windows of fixed length L. If X_t denotes the input sequence associated with time t, then this transformation was defined as

$$X_t = [y_{t-L+1}, \dots, y_t] \quad (6)$$

While the output associated with horizon h was y_{t+h}. Internally, the LSTM cell operates through gates that regulate the flow of information. Let x_t be the input at time t, h_{t-1} the previous hidden state, c_{t-1} the previous memory state, W and b the corresponding weight matrices and bias terms, $\sigma(\cdot)$ the sigmoid function, and $\tanh(\cdot)$ the hyperbolic tangent. The forget gate is defined as:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{7}$$

the input gate as

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{8}$$

the candidate state as

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{9}$$

the updated memory state as

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{10}$$

the output gate as

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{11}$$

and the hidden output as

$$h_t = o_t * \tanh(c_t) \tag{12}$$

This formulation makes it possible to represent temporal relationships that are more flexible than those captured by linear autoregressive models.

F. Configuration Selection

Since the objective of the study was to compare model families under homogeneous conditions, configuration selection was performed for each forecasting horizon using the validation set. In the case of the weekly Seasonal Naive baseline, no hyperparameter search was conducted, since a fixed seasonal period of 7 days was used for all evaluated horizons. Consequently, configuration selection was concentrated on the ARIMA, SARIMA, and LSTM models, whose final specifications were determined based on their validation performance.

For the ARIMA and SARIMA models, selection was performed over discrete combinations of non-seasonal orders and, in the case of SARIMA, seasonal orders as well. Table 1 summarizes the configurations finally selected for each forecasting horizon. This grid search procedure, complemented by the Akaike Information Criterion (AIC), constitutes a robust and widely used methodology for determining the optimal parameter structure (p,d,q) in PM2.5 time series with high volatility [2] [16]. It can be observed that, although some configurations are repeated across horizons, the optimal specification was not necessarily identical in all cases, suggesting that the most suitable temporal structure depends on the forecasting horizon considered.

In the case of LSTM, selection was carried out through a grid search over a bounded set of hyperparameters, including input window length, number of recurrent units, dropout rate, batch size, and learning rate [12]. The hyperparameter search was performed independently for each forecasting horizon using the validation set, so that the final architecture could reflect the temporal structure associated with each prediction task. The definition of these hyperparameters was aligned with standard configurations in the deep learning literature for atmospheric pollutants, using the Adaptive Moment (AdaM) optimizer and Rectified Linear Unit (ReLU) activation functions to improve training efficiency [1] [17].

TABLE 1.
Selected configurations for the ARIMA and SARIMA models according to forecasting horizon on the validation set.

Horizon (days)	ARIMA (p,d,q)	SARIMA $(p,d,q) \times (P,D,Q,s)$
1	(1,0,2)	(1,0,2)×(1,0,0,7)
7	(2,0,2)	(2,0,2)×(1,1,1,7)
14	(2,0,2)	(1,0,2)×(1,1,1,7)
21	(2,0,2)	(2,0,2)×(1,1,0,7)
28	(2,0,2)	(2,0,2)×(0,1,1,7)

Early stopping was used during training as an additional mechanism to reduce overfitting and stabilize model selection. Unlike the statistical models, the selected LSTM architecture varied mainly in window length and in the number of units. In addition, dropout rates of up to 0.2 were implemented as a regularization mechanism to prevent overfitting, which is a critical technique in recurrent models applied to complex urban environments [4]. Table 2 presents the final LSTM configurations selected for each horizon, allowing the experimental setup to be more easily reproduced.

TABLE 2.
Selected configurations for the LSTM model according to forecasting horizon on the validation set.

Horizon (days)	Lookback	Units	Dropout	Batch size	Learning rate
1	28	32	0.2	32	0.001
7	28	64	0.0	32	0.001
14	28	64	0.2	32	0.001
21	14	64	0.2	32	0.001
28	28	32	0.0	32	0.001

Taken together, the selected ARIMA, SARIMA, and LSTM configurations define the variants that were finally evaluated on the test set. The chronological partition and the use of an independent validation set for configuration selection made it possible to keep the fitting and final evaluation stages separated, following temporally consistent validation protocols to reduce the risk of methodological overfitting. This ensures a fair comparison among models of different nature under a realistic prospective scenario.

G. Updating and Evaluation Scheme

The final evaluation was carried out through a chronologically consistent protocol designed to approximate a prospective forecasting scenario [11]. For the ARIMA and SARIMA models, this procedure was implemented through a walk-forward refit scheme with an expanding window, in which the model is re-estimated at each forecast origin using all the history available up to that point and then generates the prediction for the corresponding horizon [18]. In the case of the weekly Seasonal Naive baseline, the same logic of forecast origins and horizons was maintained, although without parametric fitting, by reusing at each step the most recently observed weekly pattern. In contrast, for LSTM the operational procedure was different: for each horizon and candidate configuration, the network was previously trained on the corresponding set and then evaluated sequentially over sliding windows, without fully retraining the network at each new forecast origin [18]. Consequently, although all methods were evaluated under a prospective chronological logic and on the same target series, the comparison does not correspond to an identical operational replication across models, but rather to a consistent evaluation among forecasting strategies with different internal mechanisms.

Let T_k denote the last observed time point at forecast origin k . Then, the information available at that step can be represented as:

$$D_k = \{y_1, \dots, y_{T_k}\} \tag{13}$$

And the forecast issued at that origin for horizon h is expressed as:

$$\hat{y}_{T_k+h} = f_{k,h}(D_k) \tag{14}$$

Where $f_{k,h}(\cdot)$ represents the forecasting function associated with the model and the horizon evaluated at origin k . In ARIMA and SARIMA, this procedure involved re-estimating the model at each forecast origin using the history available up to that point and then generating the corresponding forecast. This made it possible to avoid a single static fit and to better approximate the real use of these models in operational contexts, where each new observation updates the information available for forecasting.

In the weekly Seasonal Naive baseline, the updating was implicit: at each forecast origin, the most recent weekly pattern from the series observed up to that point was reused, so the method also adapted continuously as the reference window moved forward. In other words, although there is no parametric estimation, there is still a dynamic update of the observed block used as the forecasting template.

In LSTM, the updating mechanism was implemented through configuration selection on the validation set and final evaluation on the test set, training direct models for each horizon on the regularized daily series. For each horizon, a grid of hyperparameters was defined, including window length (lookback), number of LSTM units, dropout rate, batch size, and learning rate. The configurations were compared on the validation set, and the best variant was subsequently selected for final evaluation on the test set. During sequential evaluation, the already trained model was applied over sliding windows constructed from the history available up to each forecast origin, but without complete retraining at each step.

This design sought to approximate a prospective use scenario while at the same time maintaining temporal consistency and general comparability among model families with different internal mechanisms. It also avoided artificially optimistic evaluations derived from fitting the models only once and keeping them fixed throughout the entire test period.

H. Evaluation Metrics

The predictive performance of the models was evaluated using RMSE and R^2 [1] [8]. Let y_i denote the observed value, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean of the observed values, and n the total number of evaluated predictions. The root mean square error was defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15}$$

and the coefficient of determination as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{16}$$

These metrics made it possible to compare the behavior of the different approaches across the five horizons considered, identifying both the model with the lowest error and the sensitivity of performance to increasing forecasting horizon. In this study, RMSE was considered particularly useful for penalizing larger errors, which is relevant in environmental series where episodes of higher concentration are of special interest. R^2 made it possible to evaluate the proportion of variability explained by each model compared with the observed series.

I. Environmental Engineering Contextual Considerations

Although the study did not seek to formally reconstruct critical episodes under the operational criterion of 24-hour moving concentration, the problem addressed retains direct relevance for environmental engineering. The ability to anticipate the evolution of PM2.5 at an urban station in Santiago is of interest for the analysis of pollution trends, the monitoring of periods with higher environmental burden, and the comparative evaluation of forecasting tools applicable to air quality surveillance.

Therefore, the choice of a univariate daily formulation should be understood as a first level of controlled methodological comparison. Its objective was not to build a complete multivariable environmental alert system, but rather to establish a clear and reproducible experimental basis for determining which model families best represent the temporal dynamics of the pollutant under homogeneous input and evaluation conditions.

J. Computational Environment

The experiments were implemented in Python 3.12.12. The ARIMA and SARIMA models were developed using the statsmodels 0.14.6 library, whereas the LSTM model was implemented in TensorFlow 2.19.0 using the Keras 3.10.0 API. Preprocessing and evaluation were carried out with pandas 2.2.2, NumPy 2.0.2, and scikit-learn 1.5.2. The experiments were executed on a computer equipped with an AMD Ryzen 7 250 processor, 32 GB of RAM, and an NVIDIA GeForce RTX 5060 GPU with 8 GB of memory.

III. RESULTS

These results should be interpreted as a temporally consistent comparison among model families, although the updating mechanisms differed according to each model type. Performance was analyzed quantitatively and visually across forecasting horizons of 1, 7, 14, 21, and 28 days. First, RMSE and R^2 were used to identify differences in error and explanatory capacity among the evaluated approaches.

Subsequently, this analysis was complemented with a qualitative inspection of the temporal behavior of the predictions during the winter of 2025 and with observed-versus-predicted scatter plots for the 14-day horizon, in order to examine in greater detail the ability of each model to follow the dynamics of the series and reproduce its variability during a period of greater environmental interest. It should be recalled that, although the study was conducted on calendar-day averages, the results remain useful for characterizing the evolution of PM2.5 concentrations in terms of relative magnitude and temporal behavior, especially during months in which air quality becomes more relevant for environmental surveillance.

Fig. 2 summarizes the evolution of model performance on the test set as the forecasting horizon increases, while Table 3 and Table 4 present the exact RMSE and R^2 values, respectively. In general terms, a deterioration in predictive performance is observed in all models as the forecasting horizon becomes longer, which is expected in an environmental time series forecasting problem. However, this degradation was not homogeneous across approaches.

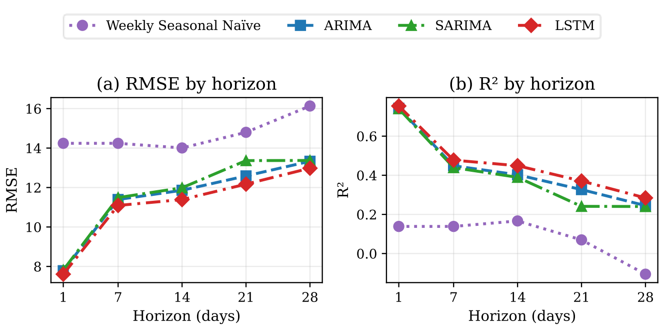

Fig. 2 Comparison of the performance of weekly Seasonal Naive, ARIMA, SARIMA, and LSTM on the test set for forecasting horizons of 1, 7, 14, 21, and 28 days, using RMSE and R^2 as evaluation metrics.

According to Table 3, LSTM achieved the lowest RMSE across the five evaluated horizons, with values of 7.61, 11.09, 11.39, 12.17, and 12.98 for 1, 7, 14, 21, and 28 days, respectively. ARIMA and SARIMA showed intermediate and relatively close performance, although systematically below that of LSTM. In contrast, the weekly Seasonal Naive baseline presented the highest errors in almost all cases, reaching an RMSE of 16.13 at the 28-day horizon. From an applied perspective, these results indicate that model choice has an important effect on forecasting accuracy, especially when the objective is to anticipate pollution trends several weeks in advance.

The same trend is observed in Table 4. LSTM achieved the highest R^2 values across all horizons considered, from 0.754 at the 1-day horizon to 0.284 at the 28-day horizon, confirming a better ability to explain the observed variability of the series in comparison with the other approaches. ARIMA and SARIMA retained positive and intermediate values, whereas weekly Seasonal Naive showed the lowest explanatory power and decreased to an R^2 of -0.106 at the 28-day horizon. Taken together, the quantitative evidence presented in Fig. 2, Table 3, and Table 4 shows that LSTM maintained a consistent advantage over the statistical models and the naive baseline throughout the entire range of evaluated forecasting horizons.

TABLE 3.
RMSE values obtained by weekly Seasonal Naive, ARIMA, SARIMA, and LSTM on the test set for forecasting horizons of 1, 7, 14, 21, and 28 days.

Model	1 day	7 days	14 days	21 days	28 days
LSTM	7.61	11.09	11.39	12.17	12.98
ARIMA	7.78	11.39	11.86	12.59	13.33
SARIMA	7.82	11.49	11.99	13.37	13.37
Weekly Seasonal Naive	14.24	14.24	14.01	14.80	16.13

Fig. 3 shows the subset of the test period corresponding to the winter of 2025, using a forecasting horizon of 14 days. This period is especially relevant from an environmental perspective, given that in Santiago PM2.5 concentrations tend to become more important during the colder months, when atmospheric conditions favor pollutant accumulation. Within this interval, ARIMA, SARIMA, and LSTM generally succeed in following the evolution of the series, whereas weekly Seasonal Naive exhibits greater variability and lower stability in response to the observed changes.

TABLE 4.
R^2 values obtained by weekly Seasonal Naive, ARIMA, SARIMA, and LSTM on the test set for forecasting horizons of 1, 7, 14, 21, and 28 days.

Model	1 day	7 days	14 days	21 days	28 days
LSTM	0.754	0.477	0.449	0.371	0.284
ARIMA	0.743	0.449	0.402	0.327	0.245
SARIMA	0.740	0.439	0.390	0.241	0.240
Weekly Seasonal Naive	0.138	0.138	0.167	0.070	-0.106

Visually, LSTM provides the smoothest and most consistent tracking of the overall trajectory during the winter 2025 period. Nevertheless, the same figure also shows that abrupt local maxima are not fully reproduced by any of the evaluated models. This point is relevant because high-concentration episodes are precisely the cases of greatest practical interest for environmental surveillance and preventive management. Therefore, the lower RMSE achieved by LSTM should be interpreted as evidence of better average forecasting behavior across the test period, rather than as a complete resolution of peak-event prediction. Under the strictly univariate daily formulation used in this study, abrupt PM2.5 increases remain difficult to anticipate, especially at an intermediate forecasting horizon such as 14 days.

Fig. 4 presents the relationship between observed and predicted values for the 14-day horizon. In the case of weekly Seasonal Naive, greater dispersion and a weaker correspondence with the reference diagonal can be observed, which is consistent with its lower quantitative performance. In practical terms, this greater dispersion reflects a lower ability to reproduce the magnitude of the observed concentrations consistently.

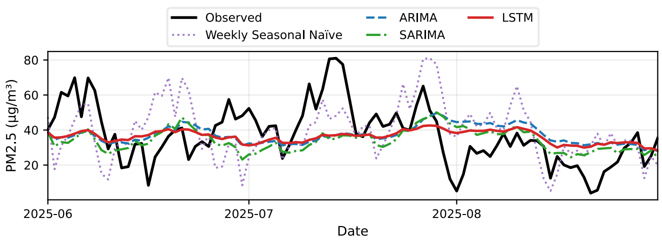

Fig. 3 Comparison between observed and predicted PM2.5 values during the winter of 2025 on the test set for a forecasting horizon of 14 days.

ARIMA, SARIMA, and LSTM show a more structured relationship between observations and predictions, although in all three cases there is a tendency to compress the dynamic range of the series, particularly at elevated PM2.5 concentrations. In other words, the predictions tend to concentrate around central values and less frequently reproduce the highest observed concentrations. This pattern is consistent with the winter 2025 temporal visualization and indicates that the main limitation is not only the average error level, but also the reduced sensitivity of the models to abrupt high-concentration events.

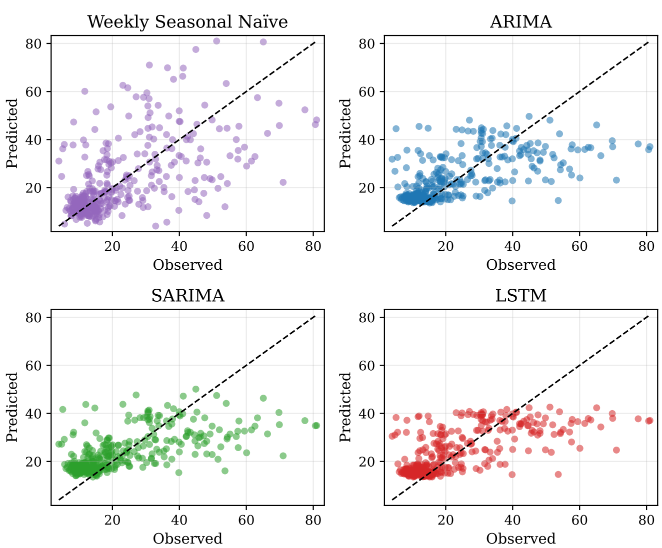

Fig. 4 Observed-versus-predicted scatter plots for weekly Seasonal Naive, ARIMA, SARIMA, and LSTM on the test set, considering a forecasting horizon of 14 days.

In visual terms, LSTM presents a more concentrated and stable cloud of points, consistent with its better overall performance in RMSE and R^2. However, the difficulty of representing the upper tail of the observed distribution remains relevant for environmental engineering applications.

Taken together, the quantitative and qualitative results show that LSTM presented the best overall performance in the evaluated scenario, whereas ARIMA and SARIMA remained competitive alternatives with intermediate behavior, and weekly Seasonal Naive exhibited the weakest fit across all analyzed horizons. Therefore, the findings reinforce the usefulness of comparing temporal modeling approaches to support the analysis of PM2.5 series in urban contexts where anticipating changes in pollutant concentration is of environmental interest.

IV. Discussion

The results obtained show that the performance of PM2.5 forecasting models depends on the forecasting horizon considered, which confirms the need for a multi-horizon evaluation in order to properly characterize the usefulness of each approach [19] [20]. A progressive deterioration in performance was observed in all evaluated models as the forecasting horizon increased, which is expected in environmental series where accumulated uncertainty grows as the anticipation period becomes longer. However, the magnitude of this degradation was not uniform across methods, revealing relevant differences in their ability to represent the temporal dynamics of the pollutant.

In this scenario, LSTM showed the best overall performance across the five evaluated horizons, both in terms of RMSE and R^2. This superiority suggests that, even under a strictly univariate formulation, a deep learning approach can capture certain temporal dependencies present in the PM2.5 series more effectively. This result is consistent with the ability of LSTM networks to model nonlinear relationships, memory effects, and complex temporal patterns that are not always represented with the same flexibility by classical linear models [21] [22]. From the perspective of environmental engineering, this finding is relevant because it suggests that more advanced temporal modeling techniques may provide concrete advantages for the predictive monitoring of atmospheric pollutants in urban contexts.

Nevertheless, the ARIMA and SARIMA models showed competitive and relatively similar performance, especially at short and intermediate horizons. This indicates that traditional statistical approaches remain valid alternatives for environmental forecasting when interpretability, lower implementation complexity, or well-established methodological frameworks are prioritized [21]. In particular, the fact that ARIMA and SARIMA maintained stable intermediate behavior suggests that an important part of the temporal structure of the series can be captured through linear dependencies and explicit seasonal components [9].

The weaker performance of the weekly Seasonal Naive baseline indicates that PM2.5 dynamics cannot be explained by direct weekly repetition alone. The series contains additional variability that requires more flexible temporal models.

The qualitative inspection of the predictions during the winter of 2025 and of the scatter plots complements this interpretation. During the winter period, the most competitive models were able to generally follow the evolution of the series, although all of them showed difficulties in fully reproducing local maxima. This point is important from an environmental perspective, since episodes of high concentrations are precisely those that tend to be of greatest relevance for air quality surveillance. In this sense, the results suggest that, although LSTM offers a better global representation of the series, the prediction of more intense episodes remains a challenge, particularly when working with a single predictor and forecasting horizons of several weeks.

Another relevant aspect is uncertainty quantification. This study evaluates point forecasts using RMSE and R^2, which is appropriate for comparing average predictive performance. However, operational air quality management may also require prediction intervals or probabilistic forecasts, especially when expected concentrations approach ranges of operational concern. Future implementations should therefore estimate not only expected PM2.5 levels, but also the uncertainty associated with each forecast horizon.

This study also presents limitations that should be considered when interpreting its results. First, the analysis is restricted to a single urban station in Santiago, so its conclusions should not be automatically generalized to other geographic contexts or to monitoring networks with different characteristics. Second, the approach is strictly univariate, so it does not incorporate exogenous variables that could strongly influence the evolution of PM2.5, such as temperature, wind, relative humidity, emissions, or topographic factors. Third, the target variable was constructed using calendar-day averages, whereas the regulatory criteria for critical episodes in Chile are defined on the basis of 24-hour moving concentrations. Therefore, the results should be understood as a comparative study of temporal forecasting rather than as a direct reconstruction of the official operational system for critical episode management.

Even with these limitations, the findings are useful as a starting point for future multivariate, multi-station, and more operationally oriented extensions for environmental surveillance applications. In particular, the study shows that a controlled methodological comparison makes it possible to identify substantive differences among model families and provides relevant evidence for the design of predictive tools applied to atmospheric pollution in Latin American urban contexts.

V. Conclusion

This study presented a controlled methodological comparison among a weekly Seasonal Naive baseline, ARIMA, SARIMA, and LSTM for daily univariate PM2.5 forecasting at an urban monitoring station in Santiago, Chile, using official SINCA data. The evaluation considered horizons of 1, 7, 14, 21, and 28 days under a chronological validation and test protocol.

The results showed that model performance varies with the forecasting horizon and that LSTM achieved the best overall performance according to the metrics considered. In particular, LSTM consistently outperformed the other approaches across the five evaluated horizons, whereas ARIMA and SARIMA maintained intermediate and competitive behavior, and the weekly Seasonal Naive baseline showed the weakest fit. These results indicate that, even under a strictly univariate formulation, the choice of forecasting model has a meaningful effect on the quality of daily PM2.5 predictions.

From an applied perspective, the study provides a reproducible benchmark for evaluating temporal modeling strategies in urban atmospheric pollution problems. The findings suggest that deep learning models can provide more accurate average forecasts than classical statistical alternatives, while also showing that abrupt high-concentration episodes remain difficult to reproduce fully. This is relevant for environmental analysis and air quality surveillance, since the most extreme concentration increases are often the events of greatest practical interest.

Future work should extend this benchmark toward multivariate and uncertainty-aware forecasting frameworks. In particular, it would be relevant to incorporate meteorological predictors, additional pollutants, emission-related variables, and information from nearby monitoring stations, as well as prediction intervals or probabilistic forecasting strategies. Future evaluations should also include criteria focused specifically on high-concentration episodes, in order to assess not only average predictive accuracy but also the ability of the models to support environmental decision-making during periods of increased pollution risk.

Acknowledgment

The authors would like to express their sincere gratitude to Gabriel Ortega for his valuable recommendations and methodological insights, which contributed to the development of this work. The authors also acknowledge the support of the Escuela de Ciencias Digitales y Tecnologías Avanzadas of Universidad Bernardo O'Higgins for providing an academic environment conducive to the advancement of this research.

References

[1] W. Ban and L. Shen, "PM2.5 Prediction Based on the CEEMDAN Algorithm and a Machine Learning Hybrid Model," Sustainability, vol. 14, no. 23, Art. no. 16128, 2022, doi: 10.3390/su142316128.

[2] L. Zhao, Z. Li, and L. Qu, "Forecasting of Beijing PM2.5 with a hybrid ARIMA model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition," Heliyon, vol. 8, no. 12, Art. no. e12239, 2022, doi: 10.1016/j.heliyon.2022.e12239.

[3] Ministry of the Environment, "Primary air quality standard for fine particulate matter (PM2.5)," [Online]. Available: https://normasaire.mma.gob.cl/nambientales/normas-de-calidad-del-aire/norma-primaria-de-calidad-del-aire-para-material-particulado-fino-mp25/. Accessed: Feb. 1, 2026.

[4] B. Peralta, T. Sepúlveda, O. Nicolis, and L. Caro, "Space-Time Prediction of PM2.5 Concentrations in Santiago de Chile Using LSTM Networks," Appl. Sci., vol. 12, no. 22, Art. no. 11317, 2022, doi: 10.3390/app122211317.

[5] M. Préndez, P. Nova, H. Romero, F. Mendes, and R. Fuentealba, "Representativeness of the particulate matter pollution assessed by an official monitoring station of air quality in Santiago, Chile: projection to human health," Environ. Geochem. Health, vol. 45, pp. 2985–3001, 2023, doi: 10.1007/s10653-022-01390-x.

[6] Ministry of the Environment, "National Air Quality Information System (SINCA)," [Online]. Available: https://sinca.mma.gob.cl. Accessed: Feb. 1, 2026.

[7] Z. Zhao et al., "Study on the prediction effect of a combined model of SARIMA and LSTM based on SSA for influenza in Shanxi Province, China," BMC Infect. Dis., vol. 23, Art. no. 71, 2023, doi: 10.1186/s12879-023-08025-1.

[8] Y. Ensafi, S. Hassanzadeh Amin, G. Zhang, and B. Shah, “Time-series forecasting of seasonal items sales using machine learning – A comparative analysis,” Int. J. Inf. Manag. Data Insights, vol. 2, no. 1, Art. no. 100058, 2022, doi: 10.1016/j.jjimei.2022.100058.

[9] M. L. Hossain, S. M. N. Shams, and S. M. Ullah, “Time-series and deep learning approaches for renewable energy forecasting in Dhaka: a comparative study of ARIMA, SARIMA, and LSTM models,” Discov. Sustain., vol. 6, Art. no. 775, 2025, doi: 10.1007/s43621-025-01733-5.

[10] Ministry of the Environment, “La Florida Station (Accredited),” National Air Quality Information System (SINCA), [Online]. Available:https://sinca.mma.gob.cl/index.php/estacion/index/id/262. Accessed: Feb. 1, 2026.

[11] E. P. Wahyuddin et al., “Improved LSTM hyperparameters alongside sentiment walk-forward validation for time series prediction,” J. Open Innov. Technol. Mark. Complex., vol. 11, no. 1, Art. no. 100458, 2025, doi: 10.1016/j.joitmc.2024.100458.

[12] D. R. Khan, A. B. Patankar, and A. Khan, “An experimental comparison of classic statistical techniques on univariate time series forecasting,” Procedia Comput. Sci., vol. 235, pp. 2730–2740, 2024, doi: 10.1016/j.procs.2024.04.257.

[13] A. Flores, H. Tito-Chura, O. Cuentas-Toledo, V. Yana-Mamani, and D. Centty-Villafuerte, “PM2.5 Time Series Imputation with Moving Averages, Smoothing, and Linear Interpolation,” Computers, vol. 13, no. 12, Art. no. 312, 2024, doi: 10.3390/computers13120312.

[14] M. Ansari and M. Alam, “An Intelligent IoT-Cloud-Based Air Pollution Forecasting Model Using Univariate Time-Series Analysis,” Arab. J. Sci. Eng., vol. 49, no. 3, pp. 3135–3162, 2024, doi: 10.1007/s13369-023-07876-9.

[15] H. Alizadegan, B. Rashidi Malki, A. Radmehr, H. Karimi, and M. Asghari Ilani, “Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction,” Energy Explor. Exploit., vol. 43, no. 1, pp. 281–301, 2025, doi: 10.1177/01445987241269496.

[16] S. Sah, B. Surendiran, R. Dhanalakshmi, and M. Yamin, “Covid-19 cases prediction using SARIMAX Model by tuning hyperparameter through grid search cross-validation approach,” Expert Syst., vol. 40, no. 5, Art. no. e13086, 2023, doi: 10.1111/exsy.13086.

[17] A. G. AbdElkader, H. ZainEldin, and M. M. Saafan, “Optimizing wind power forecasting with RNN-LSTM models through grid search cross-validation,” Sustain. Comput. Inform. Syst., vol. 45, Art. no. 101054, 2025, doi: 10.1016/j.suscom.2024.101054.

[18] M. R. Nurhambali, Y. Angraini, and A. Fitrianto, “Implementation of Long Short-Term Memory for Gold Prices Forecasting,” Malays. J. Math. Sci., vol. 18, no. 2, pp. 399–422, 2024, doi: 10.47836/mjms.18.2.11.

[19] 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. no. 107691, 2023, doi: 10.1016/j.envint.2022.107691.

[20] S. Sun, Y. Liu, Q. Li, T. Wang, and F. Chu, “Short-term multi-step wind power forecasting based on spatio-temporal correlations and transformer neural networks,” Energy Convers. Manag., vol. 283, Art. no. 116916, 2023, doi: 10.1016/j.enconman.2023.116916.

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

[22] C. Wu, R. Wang, S. Lu, J. Tian, L. Yin, L. Wang, and W. Zheng, “Time-Series Data-Driven PM2.5 Forecasting: From Theoretical Framework to Empirical Analysis,” Atmosphere, vol. 16, no. 3, Art. no. 292, 2025, doi: 10.3390/atmos16030292.