Evaluation of a Transformer-Based Model for the Temporal Forecast of Coarse Particulate Matter (PMCO) Concentrations

Luis Eduardo Mauricio-Álvarez[1], Marco Antonio Aceves-Fernandez[1,*], Jesús Carlos Pedraza-Ortega[1], and Juan Manuel Ramos-Arreguín[1]

[1] *Faculty of Engineering, Autonomous University of Queretaro, Cerro de las Campanas, Querétaro, 76010, Querétaro, México*
[*] *Corresponding author: marco.aceves@uaq.mx*

Compiled February 1, 2024

Recent advances in deep learning techniques for time series interpretation have enhanced the efficiency of environmental data analysis. One such area of focus is the forecasting of coarse particulate matter (PMCO) concentrations, known for their detrimental effects on human respiratory health. This paper employs a transformer-based neural network, renowned for its effectiveness in natural language processing and identifying complex patterns, to understand the nonlinear dynamics of PMCO concentrations. Using 2022 data from Mexico City and evaluating results at 12, 24, 48, and 72-hour forecasts, the model demonstrated a robust capability in capturing the inherent nonlinearities of PMCO. These findings set a promising foundation for the use of deep learning in air quality forecasting and can help to improve public health policies.

Keywords: Deep learning, Forecasting, Transformer, Air pollution, PMCO.

1. INTRODUCTION

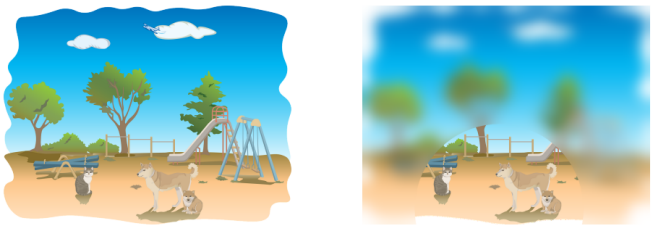
Air pollution poses a significant threat to public health and the environment globally. The high concentration of coarse particles, known as PMCO, with diameters $\geq 2.5\mu m$ and $\leq 10\mu m$ (Preiss, P., & Roos, J., 2013), has been linked to various respiratory and cardiac conditions, resulting in substantial expenditure in the health sector (European Environment Agency [EEA], 2020; Bartzis et al. 2019). These particles emanate from sources like industry, transportation, and fossil fuel burning, and their behavior and distribution patterns often display nonlinear patterns, posing significant analytical challenges (Manisalidis et al., 2020).

In this scenario, research using Artificial Neural Networks (ANNs) has gained importance. Oprea and Matei (2010) used a Feedforward Neural Network for various pollutants. Sánchez et al. (2013) compared an Elman Recurrent Neural Network and an ARIMA model with a hybrid method for SO2 emissions. Local research, such as that by Ramírez-Montañez et al. (2019) and Becerra-Rico et al. (2020), has used LSTM and GRU Neural Networks for PM10 concentrations. Barrero-González et al. (2021) demonstrated the effectiveness of an Elman Recurrent Neural Network for pollutants like O3, PM2.5, and PM10. Following these developments, the study "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case" by Wu et al. (2020) introduced the use of Transformer-based models, marking a significant advancement in time series forecasting for health-related data.

With such tools and results, in this context, using cutting-edge Artificial Neural Networks for PMCO particles, which are a growing concern in densely populated urban areas and regions with high industrial activity, such as Mexico City, is highly beneficial for the population and environment (World Health Organization [WHO], 2023).

In response to this challenge, with the goal of presenting a method to reveal PMCO particle density behavior, this study aims to apply an advanced and potentially more effective strategy using a probabilistic transformer-based neural network for the temporal prediction of PMCO concentrations (Zeng et al., 2022). We focus on analyzing data collected in Mexico City during 2022, and as it is a probabilistic model, we will consider the top 4 monitoring stations, evaluating results at 12, 24, 48, and 72-hour forecasts to provide a clearer view of the potential range of these mechanisms in capturing the nonlinear dynamics inherent in PMCO behavior (SEDEMA, 2023).

The following sections detail the background, methodology, results, and the path towards improved forecasting of air pollution.



(a) Animals in a playground (b) Our attention is drawn to animals

Fig. 1. A visual representation of an attention mechanism (Lezmi & Xu, 2023). Panel (a) features animals in a playground with uniform attention to each. Panel (b) selectively emphasizes specific animals, demonstrating the mechanism's focus.

2. BACKGROUND

A. Urban Airborne Pollution

Air quality concerns have escalated since the industrial age, with countries like Japan, the USA, and Brazil seeing pollutant reductions, contrasting with worsening conditions in less-developed nations (Health Effects Institute, 2019). Key pollutants include gases like ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2), alongside particulate matters PM2.5 and PM10 (Manisalidis et al., 2020).

Particularly, PMCO particles (2.5-10μm), Figure 2,pose significant health risks as they penetrate deep into the lungs, causing issues like irregular heartbeats and asthma (Manisalidis et al., 2020). Their environmental impacts are also notable, affecting soil nutrients, forests, and causing acid rain (EPA, 2023).

The alarming rise in air pollution-related deaths, from 1 million in 2000 to 4.2 million in 2016, underscores the urgency of addressing PMCO pollutants. In regions like the Metropolitan Zone of the Valley of Mexico, urbanization and increased vehicular and industrial activities intensify particulate matter concerns (figure 2) (Pérez-Cirera et al., 2016).

The WHO highlights the greater risks of PM2.5 particles over PM10, especially regarding mortality and cardiovascular issues (WHO, 2023). Therefore, studying and managing PMCO pollutants, bridging PM10 and PM2.5, remains crucial, particularly in urban areas like Mexico City.

B. Time Series Forecasting

Time series forecasting has emerged as a pivotal field in both scientific and business context, witnessing remarkable progress, especially with the integration of deep learning techniques alongside traditional methodologies (Wen et al., 2022). It is essential to acknowledge the divergence between the conventional models like ARIMA and the cutting-edge deep learning approaches. These advanced strategies have introduced a novel perspective in the analysis and prediction of temporal data, enabling a more intricate and adaptive understanding of the complex trends and patterns inherent in time series (Lezmi & Xu, 2023).

B.1. Probabilistic Forecasting

In time series analysis, classical methods traditionally applied to individual series within a dataset, known as "local" methods, are contrasted with modern approaches that advocate for "global" models. These global models encompass multiple series, enabling enhanced extraction of latent representations.

The current trend leans towards training global probabilistic models rather than local point forecasting models. Deep learning, particularly neural networks, is well-suited for this task due to their ability to learn from multiple time series and accurately model data uncertainty (Rasul et al., 2020).

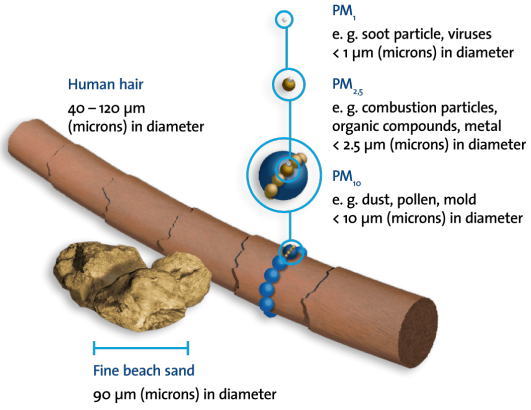


Fig. 2. Size Comparison of Particulate Matter (PM) with Human Hair and Beach Sand - A Visual Scale of PM1, PM2.5, and PM10 (EPA, 2023).

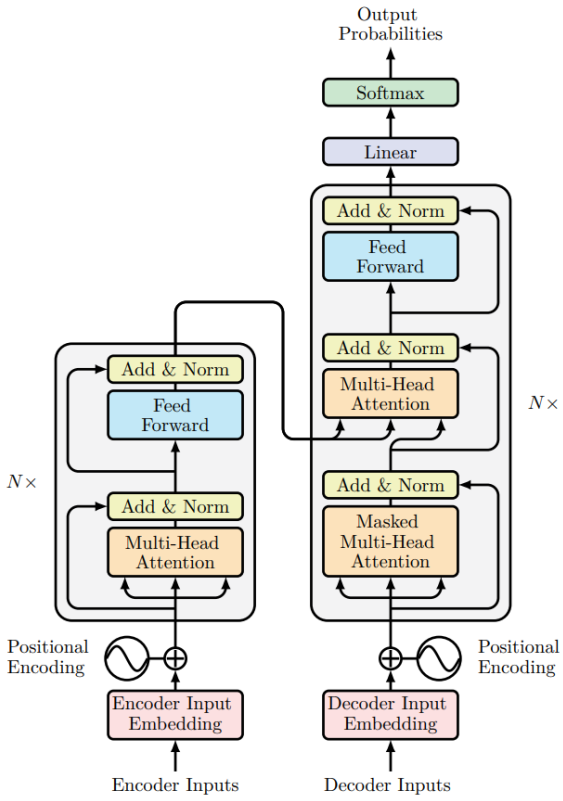


Fig. 3. The Transformer model architecture: Encoder-Decoder (Vaswani et al., 2017).

C. Transformer Neural Networks

In terms of modeling time series data, which are sequential by nature, researchers have developed models that use Recurrent Neural Networks (RNN) like LSTM or GRU, Convolutional Networks (CNN), and more recently, Transformer-based methods, which naturally fit the context of time series forecasting (figure 1) (Lezmi & Xu, 2023).

The Transformer model, introduced by Vaswani et al. (2017), is key in sequence learning tasks like machine translation, emphasizing the importance of context for understanding sequences, and its methodology can well be used for the task of univariate probabilistic forecasting (i.e., predicting the 1-D distribution of each time series individually). The Encoder-Decoder Transformer is a natural choice for forecasting as it effectively encapsulates several inductive biases (figure 3) (Lezmi & Xu, 2023).

The use of an Encoder-Decoder architecture is useful at the time of inference, where typically for some logged data we wish to forecast some prediction steps into the future. This can be considered analogous to the text generation task where, given some context, we sample the next token and pass it back into the decoder (also called "autoregressive generation").

C.1. Attention Mechanism

Attention mechanisms are the heart of Transformer models. Their operation in practical terms is like focusing on one conversation at a party while ignoring distractions. Similar to how we pay attention to specific parts in a scene (Figure 1), these mechanisms assign different weights within a sequence. This means giving higher importance to the most relevant parts for the task. This approach helps the model concentrate on key words for predicting the next, while ignoring the less important ones (Vaswani et al., 2017).

C.2. Queries, Keys, and Values

Attention mechanisms employ scaled dot-product attention, based on three components and a function:

- **Queries (Q)**: Current information under focus.
- **Keys (K)**: Labels identifying relevant parts.
- **Values (V)**: Actual data to extract based on relevance.
- **Softmax**: Used to convert weights into probabilities.

The attention equation computes weights using a scaling factor $\sqrt{d_k}$ to maintain consistent variance, aiding in training.

$$\text{Attention}(Q, K, V) = \text{softmax}_k\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad \textbf{(1)}$$

C.3. Self-Attention

Self-attention, a specific attention mechanism, uses sequences themselves to generate Q, K, and V. The input sequence X is transformed into Q, K, and V through learned parameter matrices W^Q, W^K, and W^V, as shown by equation 2-5:

$$Q = XW^Q \quad \textbf{(2)}$$

$$K = XW^K \quad \textbf{(3)}$$

$$V = XW^V \quad \textbf{(4)}$$

$$\text{Self-Attention}\,(X) = \text{softmax}\left(\frac{XW^Q\left(XW^K\right)^\top}{\sqrt{d_{\text{model}}}}\right)XW^V \quad \textbf{(5)}$$

This approach enables capturing dependencies within the sequence, crucial for sequence learning.

D. Encoder

The encoder's role is to transform an input sequence into a continuous "context" vector. The encoder is composed of N identical blocks arranged in a series, where N determines the network's depth. A higher N value signifies a more complex model with an increased number of parameters. Each encoder block, depicted on the left side of Figure 3, consists of two main components (Lezmi & Xu, 2023).

D.1. Multi-Head Attention

To enhance model flexibility and capture information from multiple perspectives, the Transformer employs multi-head attention. This involves performing the scaled dot-product attention multiple times independently and then combining the results. For h heads, this is represented as:

$$Q_i = XW_i^Q \quad \textbf{(6)}$$

$$K_i = XW_i^K \quad \text{for } i = 1, \cdots, h \quad \textbf{(7)}$$

$$V_i = XW_i^V \quad \textbf{(8)}$$

Each head computes an attention score matrix, which are then concatenated and transformed through another learned parameter matrix W^O, leading to:

$$\text{MultiHead}(X) = \text{Concat}(head_1, \ldots, head_h)W^O \quad \textbf{(9)}$$

$$\text{where } head_i = \text{Attention}Q_i, K_i, V_i) \quad \textbf{(10)}$$

Therefore, this layer measures the relevance of each element in relation to all other elements in the sequence.

D.2. Integration of Feed-Forward Layers

Following the attention process, Transformers employ feed-forward layers to apply further transformations to the data. Each encoder block within the Transformer includes a multi-head self-attention layer and a subsequent feed-forward network (Lezmi & Xu, 2023).

- The first layer applies the ReLU activation function, introducing non-linearity and computational efficiency.
- The second layer is a linear transformation, enhancing the model's ability to learn complex patterns.

These layers act as hidden layers, akin to those in classical neural networks, refining the output by combining the insights from the attention mechanism with advanced transformations.

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad \textbf{(11)}$$

Here, W_1, W_2 represent weight matrices, and b_1, b_2 are bias vectors, which are all learned during training.

The feed-forward network thus ensures the integration of the focused attention processing with powerful data transformation capabilities.

E. Decoder

In the Transformer model, the decoder is composed of a stack of identical decoder blocks, each with three components:

- **Masked Multi-Head Self-Attention Layer**: Similar to the encoder, this layer captures the relationships between elements. Inputs are processed simultaneously, with linear transformations in each head to capture different aspects of data. The decoder, however, predicts the output sequence one element at a time, based on previously generated elements. To prevent the use of future information, a padding mask is added to the input, ensuring that upcoming elements in the sequence remain hidden during training, a process known as masked forward looking (Lezmi & Xu, 2023).
- **Multi-Head Attention Layer**: This layer differs from those in the encoder block and the masked multi-head self-attention layer in the decoder. Here, the query originates from the preceding component in the decoder block, while the key and value are derived from the encoder's output. This mechanism is crucial for identifying relationships between the encoder's input and the decoder's input (Lezmi & Xu, 2023).
- **Fully Connected Feed-Forward Network**: This component is identical to the one in the encoder block (Lezmi & Xu, 2023).

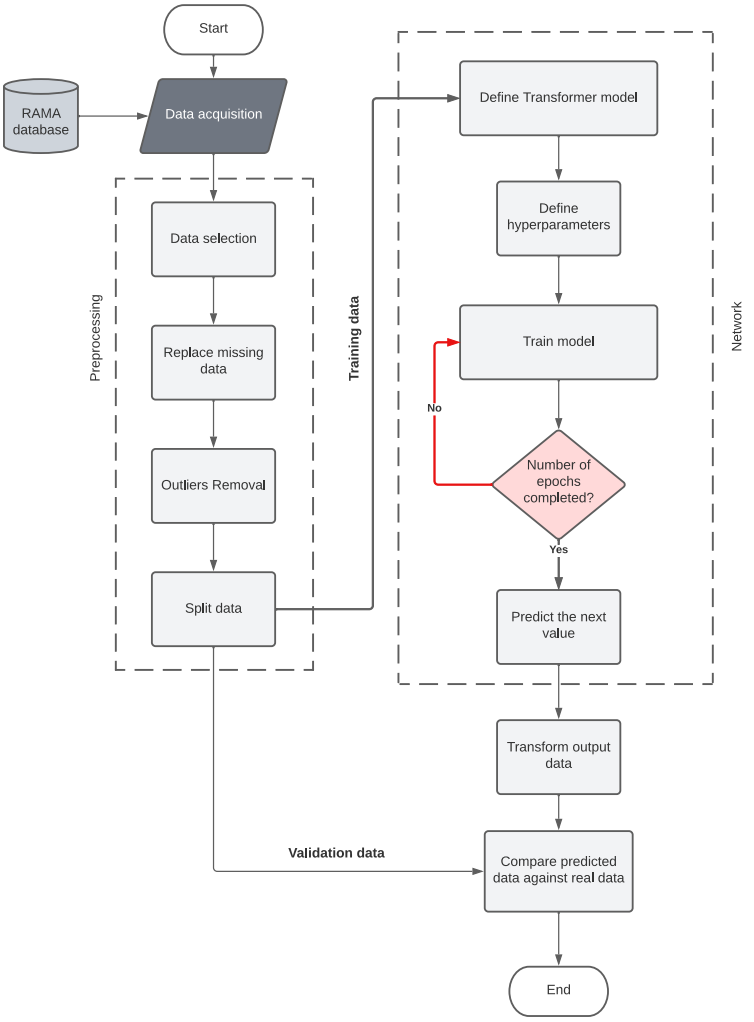


Fig. 4. Methodology flowchart.

3. MATERIALS AND METHODS

A. Materials

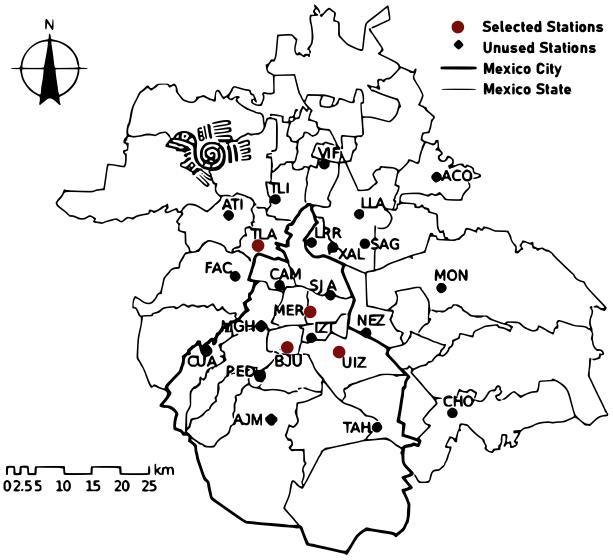


Fig. 5. Geographic location of the monitoring stations where the data was obtained (SEDEMA, 2023).

Global health protection is prioritized through continuous air quality monitoring in urban areas, as highlighted by the World Health Organization (WHO, 2023). The Atmospheric Monitoring System (SIMAT), comprising subsystems RAMA, REDMA, REDMET, and REDDA, plays a key role in this effort (SIMAT, 2023).

RAMA, a critical component of SIMAT, utilizes high-precision devices to measure pollutants such as sulfur dioxide, carbon monoxide, nitrogen dioxide, ozone, and various particulate matters across 35 Mexico City stations, supported by a laboratory for maintenance and calibration (SEDEMA, 2023).

Station selection for this study, essential for data accuracy, was based on a statistical analysis of factors like data completeness and distribution (detailed in the methodology section).

A list of selected stations in Table 1.

ID	Name	Location
BJU	Benito Juárez	Mexico City
MER	Merced	Mexico City
UIZ	UAM Iztapalapa	Mexico City
TLA	Tlalnepantla	Mexico State

Table 1. Selected Monitoring Stations for this case study.

Fig. 6. Raw data.

Table 2 summarizes the database, detailing key parameters like minimum, maximum, average values, standard deviation, and missing values for each station. Notably, "TLA" station shows a significantly higher concentration of PMCO particles and a larger standard deviation than other stations. This difference may help to ascertain the capability of this methodology to accurately forecast PMCO behavior in the presence of many missing values and outliers.

Estations	Min ($\mu g/m^3$)	Max ($\mu g/m^3$)	Mean ($\mu g/m^3$)	σ ($\mu g/m^3$)	Missing Values
BJU	1.0	146.0	13.2004	8.6353	332
MER	1.0	161.0	17.1540	10.4514	1059
UIZ	1.0	158.0	19.8663	12.8626	1247
TLA	1.0	520.0	26.8231	18.3582	1336

Table 2. Raw data statistics

Models for predicting air pollution are developed using time series data on PMCO particle concentration (measured in $\mu g/m^3$), registered hourly. Figure 6 shows 2022's raw data from selected stations. The upper chart, labeled "PMCO Data BJU 2022," does depict some missing values, potentially affecting the model's learning process. To address this, it is necessary to appropriately refine the initial data.

B. Methodology

The following methodology was carried out for the PMCO airborne particle forecasting using as training data the sites aforementioned (Figure 4):

1. **Data Acquisition:** We collect the essential raw data for prediction from the RAMA source, specifically focusing on the PMCO concentration dataset from 2022, as referenced in SEDEMA 2023.

2. **Data Selection:** For our analysis, we prioritize station databases that exhibit minimal missing data and outliers. To identify and manage outliers effectively, we use a threshold based on 3 standard deviations.

3. **Outliers Removal:** We apply the z-scores method, setting a ±3 standard deviations threshold, to efficiently eliminate outlier values.

4. **Missing Value Imputation:** To address missing data, we employ the MICE (Multiple Imputation by Chained Equations) method.

5. **Separation in Training and Validation Sets** The dataset is split into training and validation sets in a chronological order, with each month's data forming a batch. We adopt an 80/20 division, allocating 80% for training and the remaining 20% for validation, to ensure an unbiased evaluation of the model.

6. **Model Construction**

   Our model, built using the TimeSeriesTransformerForPrediction class from the Transformers library, is customized with specific parameters to optimize forecasting accuracy. Key settings include the prediction length, twice the context length for capturing temporal dynamics, and the inclusion of historical lags. The model incorporates temporal features and a unique approach to categorical features, focusing on a single static feature per PMCO time series. We balance the encoder and decoder with 4 layers each and choose a 32-dimensional model to match our data's structure. The ReLU activation function introduces necessary non-linearity, and the assumption of a Student-t distribution addresses outliers and variability.

7. **Model Training** The training process involves formatting data into batches of 256 for each of the 100 epochs per training session. Testing is conducted with smaller batches of 64. This systematic approach ensures consistent, high-quality data for optimal training and prediction accuracy.

8. **Model Optimization** During training, the AdamOptimizer monitors loss and adjusts neuron weights as needed (Diederik et al. 2014). The model employs the AdamW optimizer for its effectiveness with sparse gradients and adaptive learning rates, set with beta values of 0.9 and 0.95 to enhance convergence.

9. **Value Forecasting** After training, the model predicts future values over 12, 24, 48, and 72-hour intervals, following the established methodology.

10. **Comparison(Validation of Results)** We validate the model's performance by comparing its predictions against actual observed values using the Root-Mean-Square Error (RMSE) metric, providing a clear measure of prediction accuracy.

    Specifically, the RMSE is calculated using the following formula:

$$\mathrm{RMSE}_{\mathrm{j}}^{\mathrm{k}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}^{k} - x_{ij}^{\mathrm{real}}\right)^2}, \quad j = 1,2,3; \qquad \textbf{(12)}$$

    where x_{ij}^{k} represents the i-th predicted value of the j-th variable for the k-th prediction, and x_{ij}^{actual} is the ith real value observed.

Parameter	Value
Train Data Loader	
Batch Size	256
Number of Batches per Epoch	100
Test Data Loader	
Batch Size	64
Optimizer	AdamW
Look Back	30 days

Table 3. Hyperparameter Selection for the Model.

4. RESULTS AND DISCUSSIONS

A. Data Cleaning and Imputation

A.1. Outlier Removal

z-score imputation method was used for tackling the outlier issues, setting a threshold at ±3 standard deviation. This approach

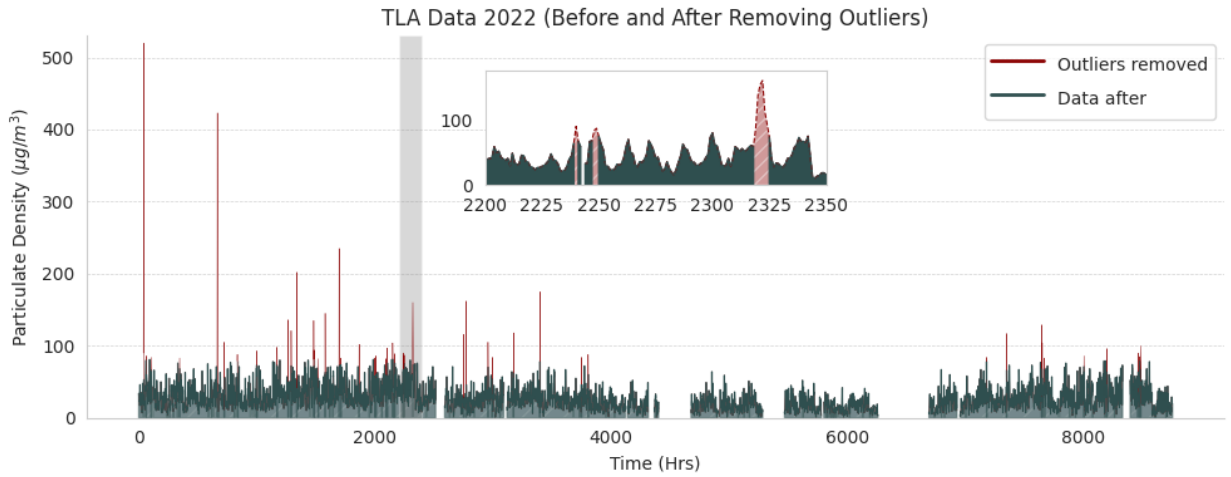


Fig. 7. TLA data 2022 with outlier removal.

proved effective for the TLA data, as it enabled us to retain a substantial 83.972603% of the dataset. This high retention rate suggests a relatively normalized distribution, allowing us to focus on the most pertinent data.

Figure 7 provides a visual representation of the data structure post-outlier removal. It depicts a more homogeneous and representative sample of the general conditions, highlighting the effectiveness of our outlier management strategy.

Additionally, we conducted a comprehensive analysis of the TLA data distribution. The Table 4 summarizes the percentage of data retained within various standard deviation ranges after the outlier removal process:

Station	1σ (%)	2σ (%)	3σ (%)
BJU	72.682 648	93.127 854	95.205 479
MER	66.141 553	84.691 781	86.929 224
UIZ	64.383 562	82.888 128	84.531 963
TLA	70.194 064	82.157 534	83.972 603

Table 4. Percentage of Data Within Different Standard Deviations after Outlier Removal

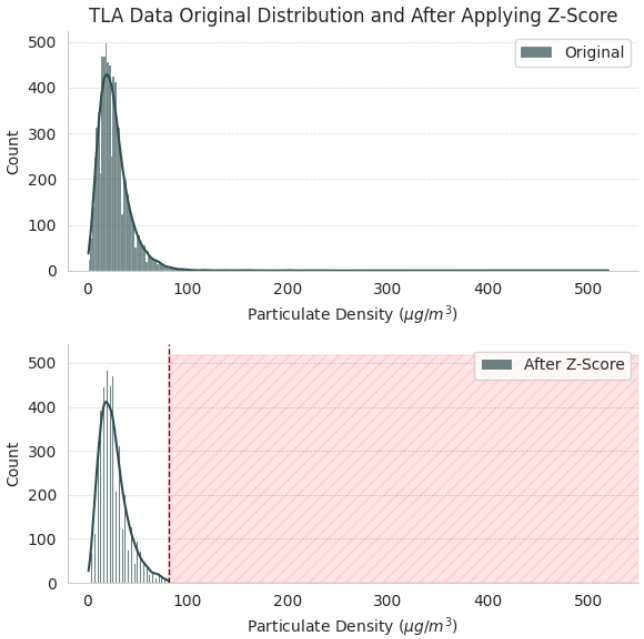
In Figure 8, we present a graph illustrating the distribution of values, particularly focusing on TLA. This visualization aids in a deeper understanding of the primary trends and behaviors of the variable, demonstrating how values are concentrated around TLA.

A.2. Missing values imputation

In the final phase of our data preparation, the dataset's remaining missing values are addressed using the MICE (Multivariate Imputation by Chained Equations) technique. This renowned method is highly effective in preserving the original data structure while robustly filling in missing information, as illustrated in Figure 9.

The Table 5 presents the final data distribution following the imputation process. This step is critical as it ensures that our subsequent analysis is based on an optimized and comprehensive dataset (Figure 9).

Fig. 8. Distribution *TLA* Data Before and After Applying Z-Score.



Estaciones	Min ($\mu g/m^3$)	Max ($\mu g/m^3$)	Mean ($\mu g/m^3$)	σ ($\mu g/m^3$)	Missing Values
BJU	1.0	39.0	12.7612	7.2498	420
MER	1.0	48.0	16.6479	9.1455	1145
UIZ	1.0	58.0	19.0268	10.5800	1355
TLA	1.0	81.0	25.9241	14.1915	1404

Table 5. Data statistics used to train the model

With the data now clean, imputed, and free from outliers, we are poised to proceed with our substantial analysis.

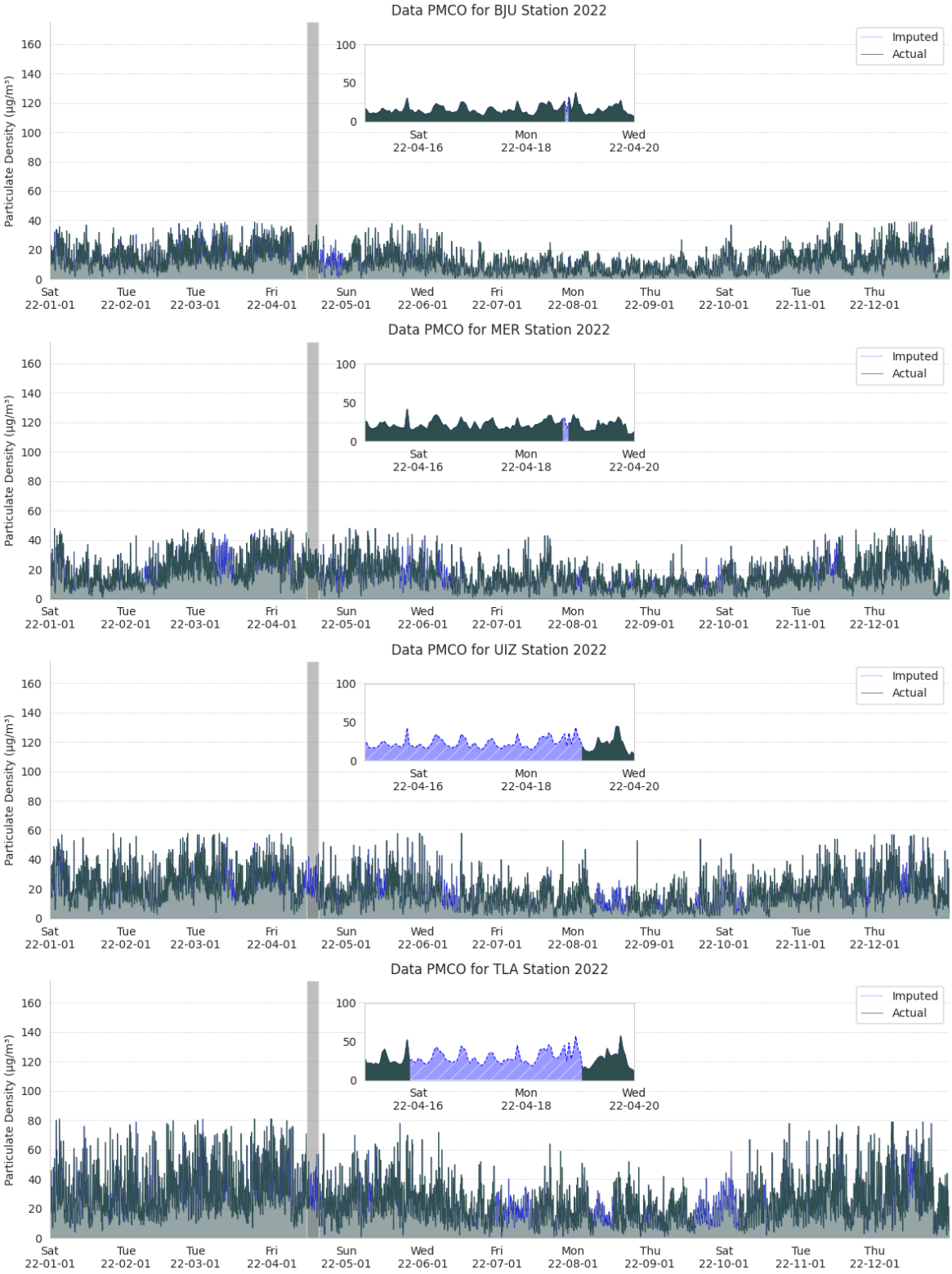


Fig. 9. Data used to train the model.

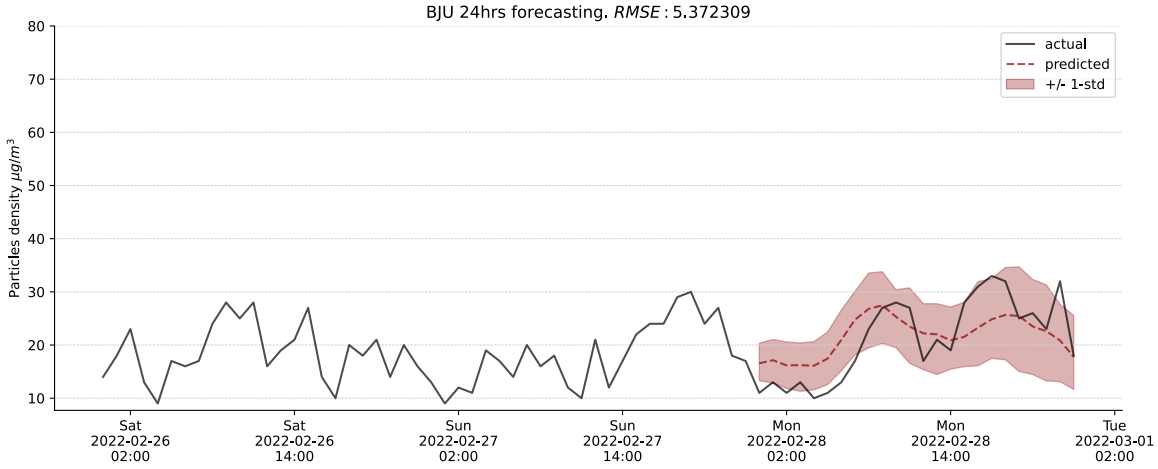


Fig. 10. Transformer-Based Model prediction of the last 24 hours of PMCO for the month of February 2022 for the BJU station.

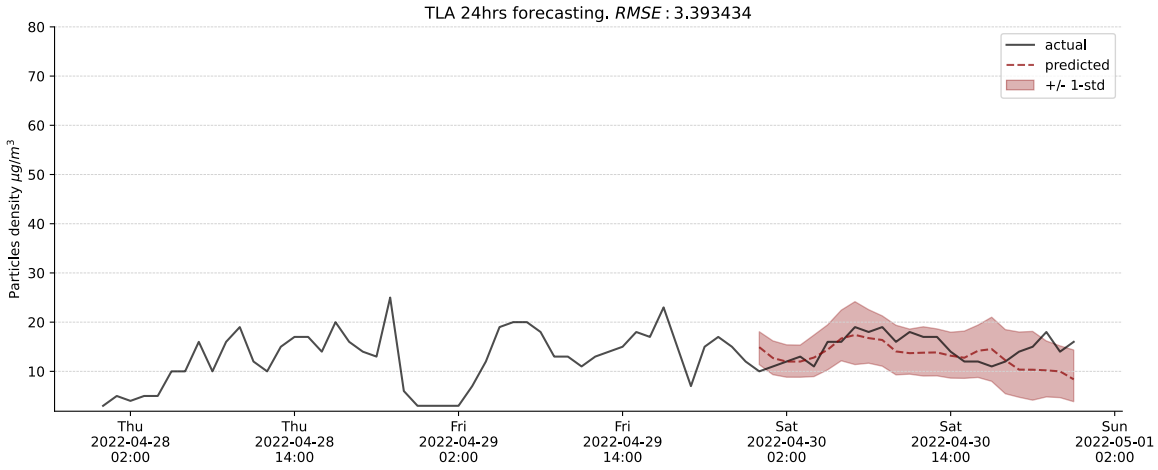


Fig. 11. Transformer-Based Model prediction of the last 24 hours of PMCO for the month of April 2022 for the TLA station.

B. Forecasting

In this stage, the model based on Transformers was tested. This model was trained with the database of the 12 months of 2022, using batches of 12, 24, 48 and 12 hours for each site mentioned in the Table 1 as indicated in the Methodology section.

In the study of forecast errors (RMSE) for the monitoring stations BJU (Benito Juárez, Mexico City), MER (Venustiano Carranza, Mexico City), TLA (Tlalnepantla de Baz, State of Mexico) and UIZ (Iztapalapa, Mexico City), several interesting trends were revealed that reflect how the specific environment and external factors impact the accuracy of forecast models.

B.1. BJU and MER: Natural Areas and Regular Vehicle Load

The stations BJU, near Parque de los Venados, and MER, close to Deportivo Venustiano Carranza, are located in areas with abundant vegetation and regular vehicle traffic. These conditions seem to contribute to greater precision in the forecasts since the majority of data present stability in their record (Figure 9). The average RMSE values for BJU fluctuate between 3.78 (24 hours) and 6.19 (72 hours), as seen in Table 6, indicating reasonable consistency in the model. Similarly, MER shows RMSE values ranging from 5.35 (24 hours) to 7.53 (72 hours), as seen in Table 7. This relative consistency could be influenced by a less variable environment and fewer industrial and vehicular emissions.

B.2. TLA and UIZ: Influence of Industrial Activity

Conversely, TLA and UIZ, located near industrial areas, show a different pattern. TLA is located next to an industrial center in Tlalnepantla de Baz, which could lead to higher particle density and, therefore, greater difficulty in prediction, due to what we mentioned previously since the much greater volatility of the data (Figure 9). This is reflected in RMSE values ranging from 5.70 (24 hours) to 9.16 (72 hours), as seen in Table 8. UIZ, situated within 2 km of significant industrial centers such as Pepsi Rojo Gómez and Grupo Collado S.A. de C.V., shows the highest RMSE values among the four stations, varying from 7.98 (24 hours) to 11.69 (72 hours), as seen in Table 9. These data suggest that proximity to intense industrial activities can significantly impact the accuracy of air quality forecasting.

B.3. Temporal Variability and Model Precision

The temporal analysis indicates that some months like May and November (Figure 12 and 13) present greater challenges in prediction, possibly due to seasonal changes in particle emissions. In contrast, months like June to October are periods with lower

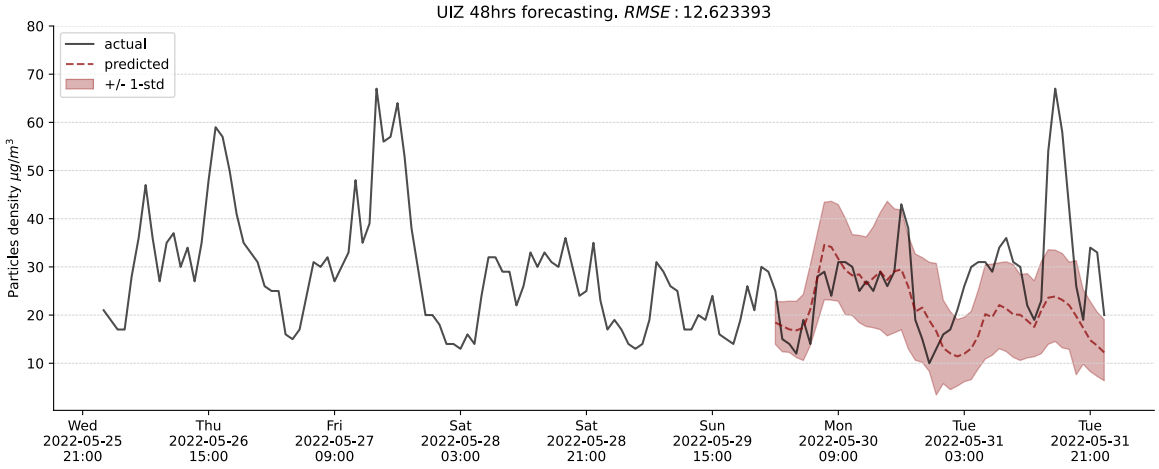


Fig. 12. Transformer-Based Model prediction of the last 48 hours of PMCO for the month of May 2022 for the UIZ station.

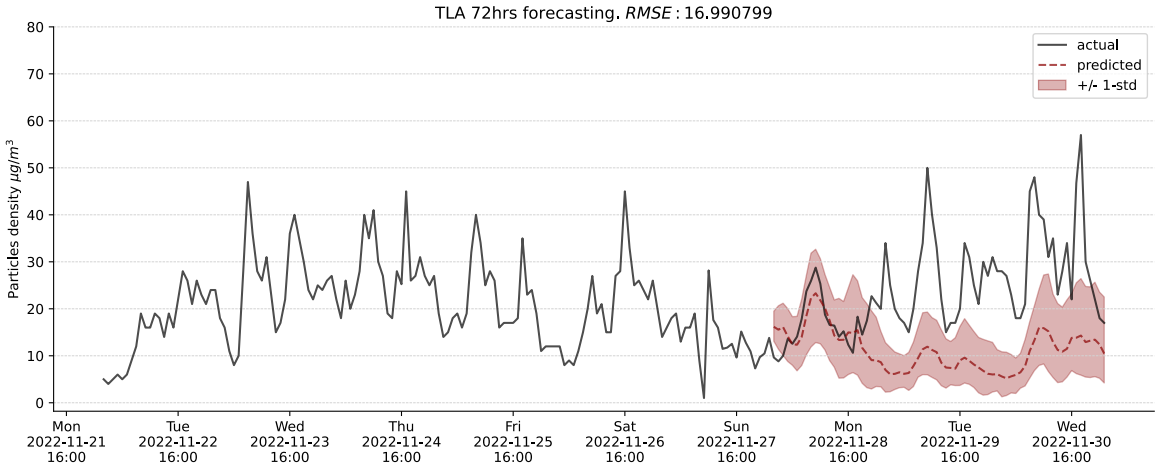


Fig. 13. Transformer-Based Model prediction of the last 72 hours of PMCO for the month of November 2022 for the TLA station.

volatility and, therefore, easier to predict. As we can see the comparison in more detail in the line graphs of Figure 15.

B.4. Predictability

The months February and April (Figure 10 and 11) are particularly noteworthy for their predictability. These months consistently show more accurate forecasts across all stations, indicating lower volatility and more stable environmental conditions. This trend is evident in the lower RMSE values, suggesting that the model performs best under the stable conditions typically present during these months.

B.5. 12-Hour Forecast Horizon

It was observed that the 12-hour forecast horizon showed significant errors in all stations, as we can see in the boxplots of each of the monitoring stations (Figure 15). This could be due to the limited capacity of the transformer models to make accurate short-term predictions or a possible insufficiency of data for this time horizon.

Forecast Horizon (hours)	12	24	48	72
Mean	6.76	3.79	5.83	6.19
Standard Deviation	3.69	1.67	2.74	3.19
Minimum	1.32	2.27	3.02	3.10
Maximum	11.66	7.10	11.13	12.71

Table 6. BJU (Benito Juárez, Mexico City)

Forecast Horizon (hours)	12	24	48	72
Mean	7.32	5.36	6.87	7.54
Standard Deviation	3.84	2.33	3.26	3.94
Minimum	3.08	2.78	3.62	3.37
Maximum	15.16	8.40	14.48	14.47

Table 7. MER (Venustiano Carranza, Mexico City)

Fig. 14. Line Charts Depicting the Monthly RMSE Trends Across Forecast Horizons at BJU, MER, TLA, and UIZ Monitoring Stations Throughout 2022.

Forecast Horizon (hours)	12	24	48	72
Mean	9.33	5.70	8.54	9.17
Standard Deviation	5.18	2.56	4.21	4.98
Minimum	3.00	3.39	3.88	4.26
Maximum	19.65	10.09	18.55	16.99

Table 8. TLA (Tlalnepantla de Baz, State of Mexico)

Forecast Horizon (hours)	12	24	48	72
Mean	13.63	7.98	11.08	11.69
Standard Deviation	8.03	3.53	5.23	5.99
Minimum	4.52	3.61	5.58	5.92
Maximum	29.12	13.49	23.29	21.89

Table 9. UIZ (Iztapalapa, Mexico City)

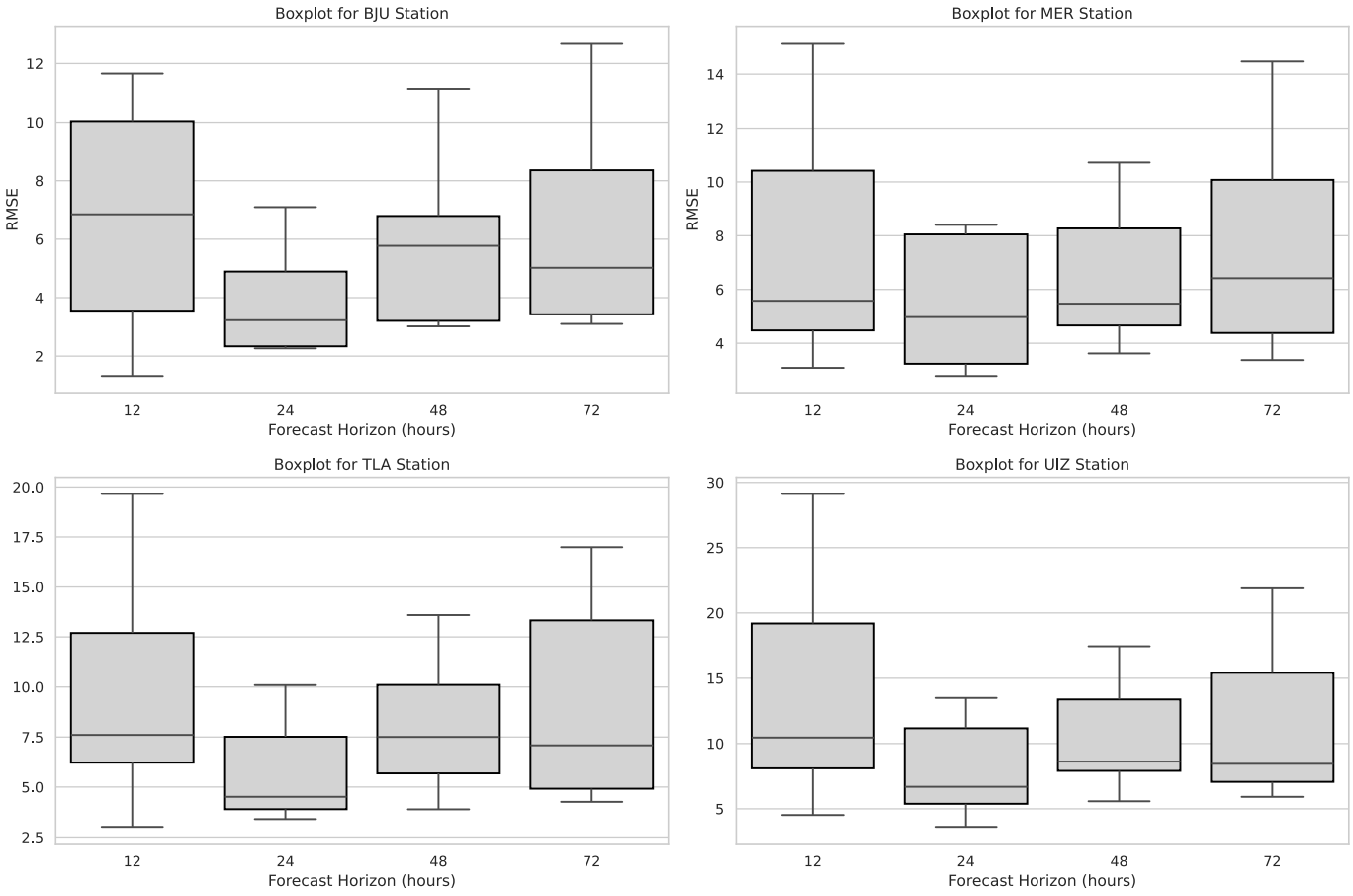


Fig. 15. Boxplot of RMSE for Transformer model predictions by site.

5. CONCLUSION

This work implemented an innovative Transformer-based model to predict PMCO concentrations in Mexico City. The results demonstrated the feasibility of modeling highly non-linear environmental behaviors with moderate accuracy using this approach. Multiple tests were conducted to assess the model's robustness over different time periods, including predictions up to 72 hours ahead. Additionally, predictions were segmented monthly, highlighting the impact of missing data and the standard deviation on PMCO predictions. Future work could explore alternative data pre-processing methods or ensemble techniques to enhance the outcomes.

REFERENCES

1. Bartzis, J. G., Kalimeri, K. K., & Sakellaris, I. A. (2019). Environmental data treatment to support exposure studies: The statistical behavior for NO2, O3, PM10, and PM2.5 air concentrations in Europe. *Environmental Research*. https://doi.org/10.1016/j.envres.2019.108864
2. Barrero-González, D., Ramírez-Montañez, J. A., Aceves-Fernández, M. A., & Ramos-Arreguín, J. M. (2021). Capability of an Elman recurrent neural network for predicting the non-linear behavior of airborne pollutants. *Earth Science Informatics*. https://doi.org/10.1007/s12145-021-00707-1
3. Becerra-Rico, J., Aceves-Fernández, M. A., Esquivel-Escalante, K., & Pedraza-Ortega, J. C. (2020). Airborne particle pollution predictive model using Gated Recurrent Unit (GRU) deep neural networks. *Earth Science Informatics, 13*(3), 821-834. https://doi.org/10.1007/s12145-020-00462-9
4. Diederik, P., et al. (2014). A method for stochastic optimization. *3rd International Conference for Learning Representations, San Diego*. arxiv:1412.6980
5. European Environment Agency. (2020). What is particulate matter and what are its effects on human health?. https://www.eea.europa.eu/help/faq/what-is-particulate-matter-and
6. EPA. (2023). Health and environmental effects of Particulate Matter (PM). https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm
7. Health Effects Institute. (2019). State of Global Air 2019. Special Report. *Health Effects Institute, Boston*.
8. Probabilistic time series forecasting with transformers. (2021). Retrieved from Huggingface.co website: https://huggingface.co/blog/time-series-transformers
9. Lezmi, E., & Xu, J. (2023). Time Series Forecasting with Transformer Models and Application to Asset Management. *Amundi Institute*.
10. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health, 8*, 14. https://doi.org/10.3389/fpubh.2020.00014
11. Oprea, M., & Matei, A. (2010). The Neural Network-Based forecasting in environmental systems. *WSEAS Transactions on Systems and Control, 5*(12), 893-901.
12. Pérez-Cirera, V., Schmelkes, E., López-Corona, O., Carrera, F., García-Teruel, A. P., & Teruel, G. (2016). Ingreso y calidad del aire en ciudades: ¿Existe una curva de Kuznets para las emisiones del transporte en la Zona Metropolitana del Valle de Mexico? *El trimestre económico, 85*(340), 745–764.

13. Preiss, P., & Roos, J. (2013). Global characterization factors for damage to human health due to particulate matter – based on the TM5-FASST model. *University of Stuttgart*. https://lc-impact.eu/doc/deliverables/fine_particular_matter.pdf
14. Ramírez Montañez, J. A., Aceves Fernández, M. A., Tovar Arriaga, S., Ramos Arreguín, J. M., & Salini Calderon, G. A. (2019). Evaluation of a Recurrent Neural Network LSTM for the detection of exceedances of particles PM10. *Proceedings of the 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 1-6. https://doi.org/10.1109/ICEEE.2019.8884516
15. Rasul, K., Sheikh, A.-S., Schuster, I., Bergmann, U., & Vollgraf, R. (2020). Multivariate probabilistic time series forecasting via conditioned normalizing flows. http://arxiv.org/abs/2002.06103
16. Sánchez, A. B., Ordóñez, C., Lasheras, F. S., de Cos Juez, F. J., & Roca-Pardiñas, J. (2013). Forecasting SO2 pollution incidents by means of Elman Artificial Neural Networks and ARIMA models. *Abstract and Applied Analysis, 2013*, Article 238259. https://doi.org/10.1155/2013/238259
17. SEDEMA. (2023). Atmospheric Monitoring Directorate. http://www.aire.cdmx.gob.mx/default.php?opc='ZaBhnmI='
18. Dirección de Monitoreo Atmosférico. (2023). http://www.aire.cdmx.gob.mx/aire/default.php
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. http://arxiv.org/abs/1706.03762
20. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2022). Transformers in time series: A survey. http://arxiv.org/abs/2202.07125
21. Organización Mundial de la Salud (WHO). (2023). Air pollution. https://www.who.int/health-topics/air-pollution#tab=tab_1
22. Wu, N., Green, B., Ben, X., & O'Banion, S. (2020). Deep Transformer models for time series forecasting: The influenza prevalence case. http://arxiv.org/abs/2001.08317
23. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2022). Are transformers effective for time series forecasting? https://arxiv.org/abs/2205.13504

DECLARATIONS

- **Funding:** No specific funding was received for conducting this study.

- **Conflict of Interest/Competing Interests:** The authors declare that they have no competing interests.

- **Ethics Approval and Consent to Participate:** Not applicable as this study did not involve human participants, human data or human tissue.

- **Consent for Publication:** Not applicable as this manuscript does not contain any individual person's data.

- **Data Availability:** The data supporting the findings of this study are openly available from the RAMA source, focusing on the PMCO concentration dataset from 2022 [Link]. Further information about the data can be provided upon request.

- **Materials Availability:** Not applicable as this study did not generate new materials.

- **Code Availability:** Due to confidentiality agreements and the proprietary nature of the computational code used in this study, the code is not publicly available. Specific inquiries about the code and its functionalities can be directed to the corresponding author.

- **Authors' Contributions:** Luis Eduardo Mauricio-Álvarez was involved in data collection and analysis. Marco Antonio Aceves-Fernandez (Corresponding Author) contributed to the study conception and design, and was responsible for the drafting of the manuscript and revisions. Jesús Carlos Pedraza-Ortega and Juan Manuel Ramos-Arreguín provided critical feedback and helped shape the research, analysis and manuscript. All authors read and approved the final manuscript.