International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 12 - Issue 07 || July 2026 || PP. 11-34

Ozone Concentrations in Mexico City using Machine Learning with Multiple Time Series

M. Sc. Zenteno Jiménez José Roberto
Geophysical Engineering, National Polytechnic Institute, Mexico City
ESIA- Ticóman Unit Mayor Gustavo A. Madero
Email: jzenteno@ipn.mx

Abstract: Tropospheric ozone (O3) pollution in Mexico City represents a critical and persistent environmental challenge, characterized by nonlinear dynamics and a strong dependence on meteorological factors and chemical precursors. This study proposes a comprehensive methodological framework to analyze and predict O3 concentrations, overcoming the limitations of traditional univariate models. The methodology is structured in three phases. First, a Vector Autoregressive (VAR) model is applied along with Impulse-Response Functions (IRFs) to examine the stochastic interdependence and the system's response to shocks in the variables. Second, Machine Learning techniques, specifically Randomization, are implemented. Forest and Gradient Boosting (GBoosting) integrates key exogenous variables such as nitrogen dioxide (NO2) and daily temperature to capture complex nonlinear patterns. Finally, long-term sequential memory capacity is evaluated using a Deep Learning architecture based on Gated Recurrent Units (GRUs). Experimental results indicate that [mention your main finding here, e.g., the hybrid model or GBoosting reduced the mean squared error by X compared to the baseline models]. The analysis confirms that the inclusion of NO2 and temperature as exogenous predictors is crucial for model accuracy. This study demonstrates that combining econometrics and neural networks offers a robust tool for environmental monitoring and public health decision-making.
Keywords: Ozone (O3), Time Series, VAR, Random Forest , Gradient Boosting , GRU, Mexico City.

Introduction

We begin with a hybrid and robust approach that covers three crucial fronts in time series analysis: statistical interdependence (VAR/IRF), nonlinear predictive capacity (Random Forest / GBoosting) and sequential deep learning (GRU).

Given the stochastic and highly nonlinear nature of air pollutants in the Valley of Mexico basin, this study adopts a comprehensive methodological approach that transcends traditional univariate models. To accurately capture the dynamics of ozone (O3) concentrations, a hybrid architecture is implemented that combines time series econometrics and ensemble machine learning. Machine Learning) and Deep Learning. Initially, to understand the linear interdependence and temporal causality between variables, the Vector Autoregressive (VAR) model is used. This technique allows meteorological and chemical variables to be treated as endogenous within the system.

Complementarily, Impulse-Response Functions (IRFs) are applied. An IRF describes how a system variable reacts to an unexpected "shock" in another variable, and how long that effect takes to dissipate. This is done to isolate and quantify the magnitude and persistence of shocks in the system, allowing visualization of how unexpected perturbations in precursors affect O3 levels over time. However, the limitations of linear models in capturing complex patterns motivate the incorporation of Machine Learning algorithms. Random Forest, a decision tree-based assembly method, is specifically designed to integrate key exogenous variables such as nitrogen dioxide (NO2) and daily temperature. This algorithm stands out for its ability to model nonlinear interactions between these precursors and ozone without the risk of overfitting. Gradient is also implemented. Boosting (GBoosting), a technique that sequentially optimizes the loss function, improving predictive accuracy in scenarios where residuals from previous models still contain exploitable structural information.

Finally, to address the long-term temporal dependencies inherent in air quality time series, a neural network based on Gated Recurrent Units (GRUs) is incorporated. Unlike standard recurrent networks, GRUs efficiently manage information flow through gating mechanisms, which is crucial for retaining relevant historical pollution patterns and filtering out short-term noise, thereby optimizing prediction over extended time horizons.

Vector Autoregressive (VAR) technique. Unlike linear regression, a VAR model does not assume a single dependent variable; it treats the three time series (O3, NO2, and Temp) as **endogenous**, influencing each other.

A VAR model predicts the value of each variable at a given time based on its own past values and the past values of the other variables. It is especially useful when relationships between variables are transmitted over several periods, correcting for endogenous problems (simultaneity between variables) and showing the relationship between stationary variables.

Lag Order Selection

In other words, how many days back will be observed to predict tomorrow's data? This is done using information criteria such as AIC (Akaike) or BIC (Bayesian). In pollution, lags of 1 to 7 days are usually acceptable.

Integrated Methodological Flowchart

The analysis process is structured in sequential and parallel stages to ensure the predictive robustness and explanatory capacity of the system. Daily concentration data from January 1 to July 1, 2025, daily maximum concentrations up to August 4, and daily maximum O3 concentration data for all of 2025 were used.

Stage	**Variables**	**Processes**
Raw Data	O3 Concentration	Data cleaning and validation.
	NO2 concentration	Data cleaning and validation.
	Daily Temperature	Data cleaning and validation.
Engineering with the Characteristics	Lagged Variables	Creation of optimal backlogs.
	Exogenous Variables	Identification of exogenous variables (NO2, Temp).

Resulting equations of the model for a lag

$$O_{3t} = \alpha_1 + \beta_{11}O_{3,t-1} + \beta_{12}NO_{2,t-1} + \beta_{13}T_{t-1} + \epsilon_{1,t}$$
$$NO_{2t} = \alpha_2 + \beta_{21}O_{3,t-1} + \beta_{22}NO_{2,t-1} + \beta_{23}T_{t-1} + \epsilon_{2,t}$$
$$T_t = \alpha_3 + \beta_{31}O_{3,t-1} + \beta_{32}NO_{2,t-1} + \beta_{33}T_{t-1} + \epsilon_{3,t}$$

Where the alphas are the intercepts, the beta coefficients are those that measure the impact of iajy, and epsilon is the white noise, which will be resolved by matrix representation

$$Y_t = A + BY_{t-1} + \epsilon_t$$

With

$$A = \begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix} \text{and} B = \begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13}\\ \beta_{21} & \beta_{22} & \beta_{23}\\ \beta_{31} & \beta_{32} & \beta_{33}\end{bmatrix}$$

Where the values of matrix A and B are calculated using the Least Squares (LS) estimation technique. To calculate the IRF, a VAR model is transformed into its **Infinity Moving Average (VMA) representation**. The vector would then be

$$y_t = \begin{bmatrix}O_{3t}\\ NO2t\\ Tt\end{bmatrix} \text{and} y_t = c + By_{t-1} + u_t$$

Where B is the matrix with the coefficients that relate each variable to its past and that of the others, and the vector u is the vector of errors or shocks, the coefficients of matrix **B** are the values obtained by **Least Squares (OLS)** for each equation of the system. In a VAR(1) model, each variable is a function of the values of yesterday (t-1) of all the variables of the system.

Where c is the expected value or baseline level of the variables when the past effect is zero. If the system is stable, the vector c determines the long-term average of the concentrations of ozone, NO2, and temperature. That is, where the pollution levels return to after the effect of a shock disappears.

The vector ut represents the **shocks** . These are movements in the variables that the model **cannot explain** with the studied past data of the time series; in our case **NO2,** a sudden fire or road blockage that generated unusual traffic; in **Temperature** , an unexpected heat wave or a cold front; and in **O3,** a day with solar radiation, which must be included, otherwise not.

GRU

GRU is an improved version of standard RNNs. The main difference lies in the RNN's compatibility with hidden state activation: a mechanism that determines when the hidden state should be updated and reset. GRU can also be considered a variant of LSTM, and both have a similar design and offer excellent predictions. In GRU, only two gates are involved in each time step: the reset gate and the update gate.

The mathematical structure of this analysis uses a given input vector, such as this, for either daily maximums or daily concentrations.

$$\boldsymbol{x_t} = \begin{bmatrix} \boldsymbol{MAXO3} \\ \boldsymbol{MAXNO2} \\ \boldsymbol{MAXTemp} \end{bmatrix}$$

The two reset gates and the update gate are used; the Zt gate is the one that updates, determining how much information $h_{t-1}$from the previous state should be passed to the future.

$$z_t = \sigma(W_z[h_{t-1}, \boldsymbol{x_t}] + b_z)$$

Now, the reset component: how much information from the past is forgotten before calculating the new memory content?

$$r_t = \sigma(W_r[h_{t-1}, \boldsymbol{x_t}] + b_r)$$

Candidate memory content, used to clear past data relevant to the current input

$$\overline{h_t} = tanh(W[r_t(.)h_{t-1}, \boldsymbol{x_t}] + b)$$

The final hidden state (step h_t) is the final combination between the previous state and the moderated candidate.z_t

$$h_t = (1 - z_t)(.)h_{t-1} + z_t(.)\overline{h_t}$$

sigmoid function between (0 to 1), (.) Hadamard multiplication , with W and b weights and bias that the model learns during training and thus generates the output as

$$\hat{y}_{t+1} = W_{out} * h_t + b_{out}$$

Extreme Gradient Boosting

XGBoost is a supervised machine learning method for classification and regression; it is an abbreviation of the English words "extreme gradient". "Boosting" (extreme gradient boosting). This method is based on decision trees and represents an improvement over other methods, such as random forest and gradient boosting. It works well with large and complex datasets by using various optimization methods.

To fit a training dataset using XGBoost , an initial prediction is made. Residuals are calculated based on the predicted and observed values. A decision tree is created from the residuals using a similarity score. The similarity of the data for a leaf is calculated, as well as the similarity gain for subsequent splits. The gains are compared to determine a feature and a threshold for a node. The output value for each leaf is also calculated using the residuals. For classification, values are typically calculated using the odds and probabilities log. The output of the tree becomes the new residual for the dataset, which is used to build another tree. This process is repeated until the residuals stop decreasing, or until the specified number of times. Each subsequent tree learns from the previous trees and is not assigned the same weight, unlike how Random Forest works.

To use this model for prediction, the output of each tree multiplied by a learning rate is added to the initial prediction to arrive at a final value or ranking.

XGBoost uses the following parameters and methods to optimize the algorithm and provide better results:
Regularization: A regularization parameter (lambda) is used when calculating similarity scores to reduce sensitivity to individual data and avoid overfitting.

Pruning: A tree complexity parameter (gamma) is selected to compare gains. The branch where the gain is less than the gamma value is pruned. This avoids over-tightening by cutting unnecessary branches and reducing tree depth.

previous O3 predictions . It calculates the gradients (gi) and Hessian gradients (hi) for each daily reading. It searches the variables (NO2 , Temperature, etc.) to determine the threshold that best separates these errors. It assigns a weight (wj) to each branch to correct the error of the previous model, using "sliding windows" (lags). Now the goal is not only to minimize error, but also to control the complexity of the model to avoid overfitting; thus, the objective function is:

$$Objt = \sum_{i=1}^{n} L\left(y_i \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Where $y_i$are the values of the time series, $\hat{y}_i^{(t-1)}$cumulative prediction of the previous tree, $\Omega(f_t)$regulation term and $f_t(x_i)$the new tree being trained, now expanding the previous objective function

$$Objt = \sum_{i=1}^{n} L\left(y_i \hat{y}_i^{(t-1)} + g_i * f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right) + \Omega(f_t)$$

With the tree structure

$$\omega_j = \frac{\sum i \epsilon I j g_i}{\sum i \epsilon I j h_i + \lambda}$$

Where Ij is the set of time series data that fall within j, and lambda is a regularization parameter to avoid very large weights. For NO2 concentrations and temperature, the model seeks the cutoff point that maximizes the gain.

$$Gain = \frac{1}{2}\left[\frac{(\sum g_L)^2}{\sum h_L + \lambda} + \frac{(\sum g_R)^2}{\sum h_R + \lambda} - \frac{(\sum g_L + \sum g_R)^2}{\sum h_L + \sum h_R + \lambda}\right] - \gamma$$

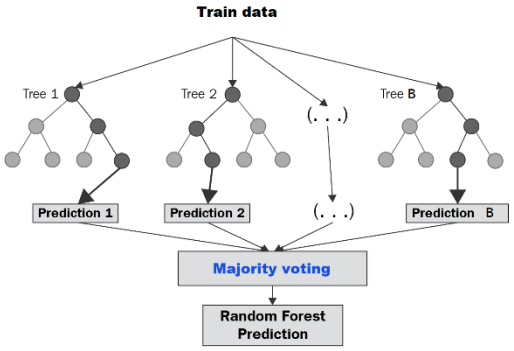
With gamma as the threshold

Random Forest

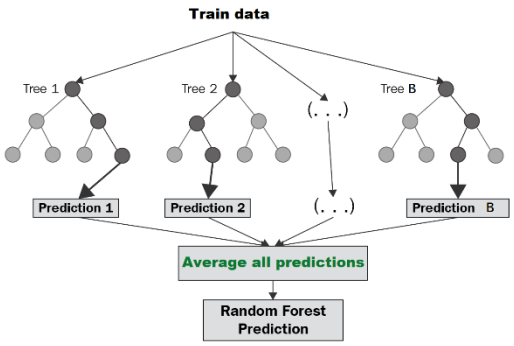
Random forest is a commonly used machine learning algorithm that combines the results of multiple decision trees to arrive at a single result. The random forest model consists of several decision trees; the decision nodes in the tree serve as a means of splitting the data. Decision trees seek to find the best split for subsets of data and are typically trained using the classification and regression tree algorithm. Metrics such as information gain or mean squared error (MSE) can be used to evaluate the quality of the split.

Decision trees are common supervised learning algorithms; they also have problems, such as bias and overfitting, but when several decision trees form an ensemble in the random forest algorithm, they predict more accurate outcomes, especially when the individual decision trees are not correlated with each other.

Para clasificación



Para regresión



Source: https://fhernanb.github.io/libro_mod_pred/rand-forests.html

Mathematically, Random Forest resides in Statistics and Averages, thus it is as follows

$$y_{RF} = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

With B being the number of trees and $f_b(x)$individual prediction of tree b for a specific day, the RF seeks to minimize impurity in each node, for regression (predicting the value of O3) the metric is variance reduction.

For a node S, the algorithm searches for the variable and the threshold (t) that maximize this reduction $\Delta V = V(s) - \left(\frac{n_{izq}}{n} V(Sizq) + \frac{n_{der}}{n} V(Sder)\right)$where V(s) is the variance of the ozone values in this case and

$$V(s) = \frac{1}{n}\sum_{i \in S}(y_i - \bar{y})^2$$

The tree chooses the cut that makes the O3 values on the right similar to those on the left.

Sequential steps to perform.

Step I: Modeling and Structure

1. **Application of VAR (Vector Autoregressive):** The linear interdependence between O3, NO2 and Temperature is modeled.
2. **IRF Calculation (Impulse-Response Functions):** Analysis of how a *shock* (impulse) in NO2 or Temperature affects O3 in future periods.
3. **Outcome:** Understanding **dynamic structure and causality** .

Step II: Predictive Machine Learning Modeling (Ensemble)

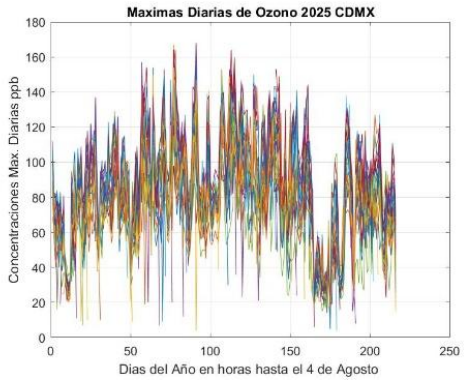
1. **Data Division:** Training, Validation and Testing.
2. **Random Training Forest :** The model is fitted using O3 lags and only the time series of daily ozone maximums from 2025
3. **GBoosting Training (Gradient) Boosting):** Sequential optimization of the model to reduce residual errors.
4. **Output:** Evaluation of the **importance of features** and nonlinear prediction.

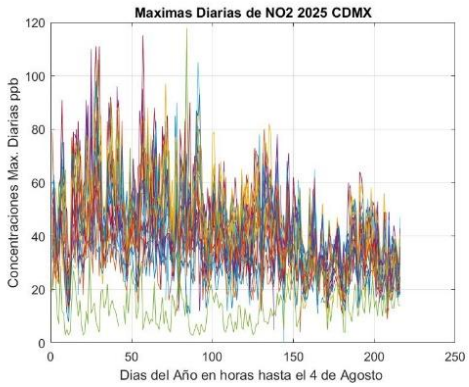
Step III: Predictive Modeling of Deep Learning (Sequential)

1. **Data Redesign:** Time Sequence Format (samples, time steps, features).
2. **GRU (Gated Recurrent Units) Training:** The network learns long-term temporal dependencies of sequential data.
3. **Output: Highly accurate** prediction of future sequences.
4.

Stage	Results	Bottom line
Performance Evaluation	Comparison of Metrics: RMSE, MAE, R^2, etc.	Identification of the model with the greatest predictive power.
Conclusions and Discussion	Integration of VAR/IRF findings, Random Forest and GRU.	**O3 forecast** and recommendations for environmental policies.

Results, using the Maximum Daily Concentration values of NO2, TEMP and O3
Graphed data





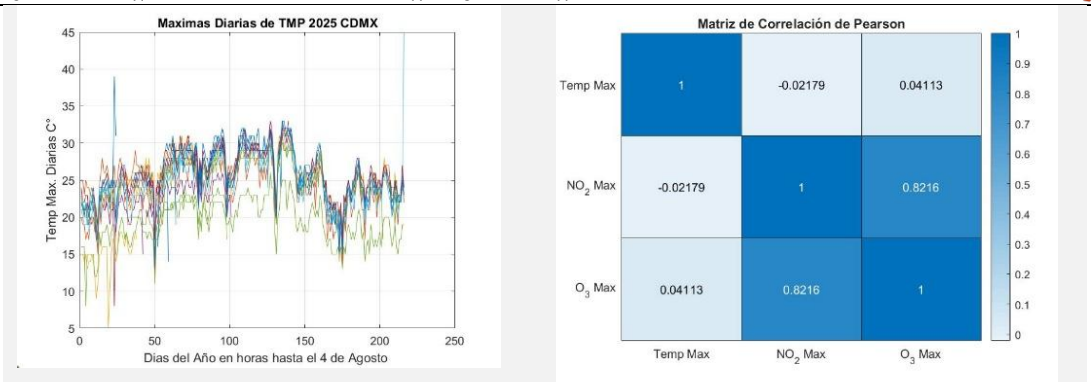


Figure 1

Results matrices

Correlation Matrix (R):	p-value matrix (P):
1.0000 -0.0218 0.0411 -0.0218 1.0000 0.8216 0.0411 0.8216 1.0000	1.0000 0.7502 0.5477 0.7502 1.0000 0.0000 0.5477 0.0000 1.0000

Validations by graphing and creating histograms

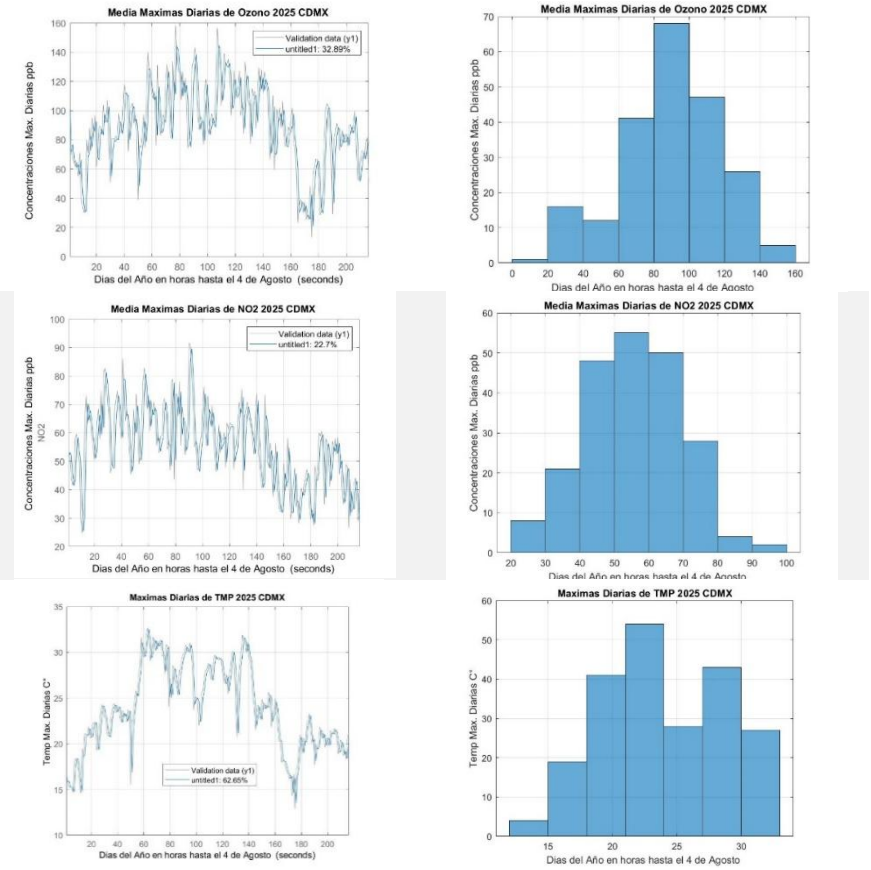


Figure 2

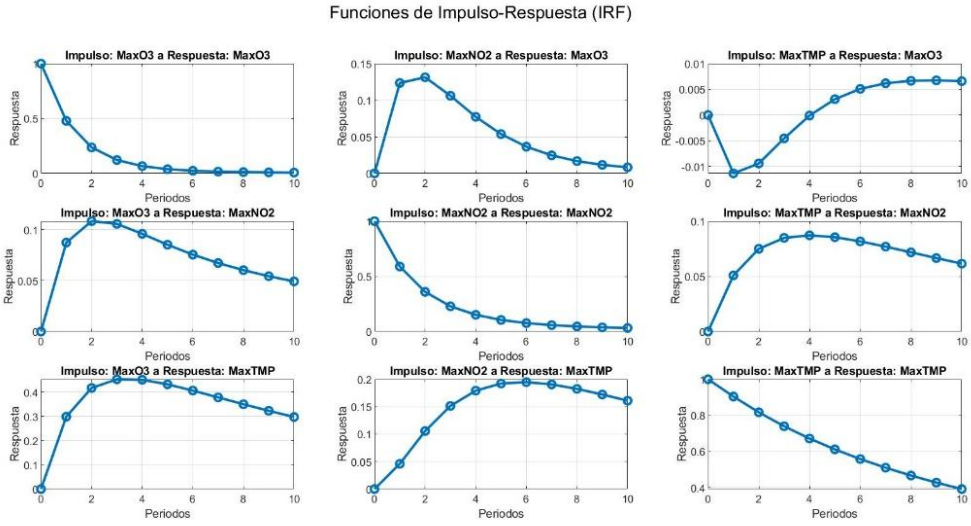


Figure 3

Impulse-response graphs, to analyze the dynamics between temperature, NO2 and O3. The shapes of the curves show how a change in one variable (the "impulse") affects the others over time.

Impulse-Response Functions (IRFs) show how an unexpected change in one variable affects the others. Now, with the forecasting function, it's possible to predict how the three time series will behave in the future, based on their historical behavior.

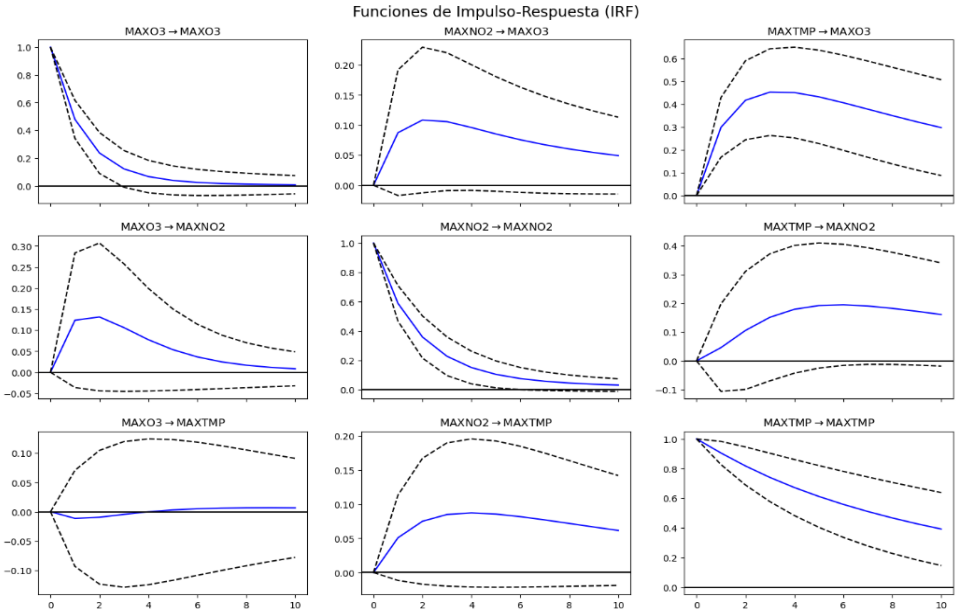


Figure 4

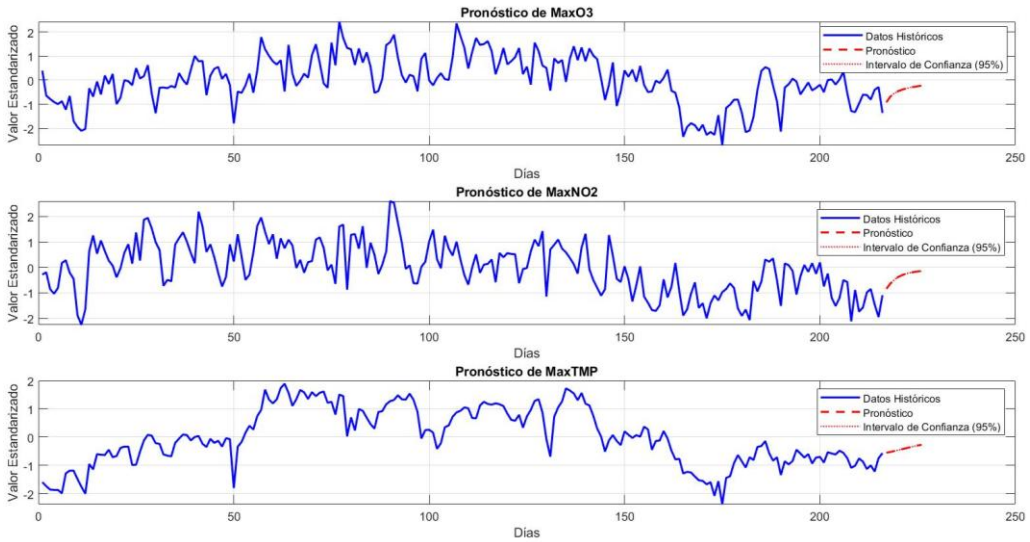


Figure 5
Forecasting Function
VAR Results

Data loaded. Observations: 216, Variables: 3

Information criteria for the delay order: VAR Order Selection (* highlights the minimums)

```
==================
============================
AIC BIC FPE HQIC
------------------------------------------------
0 -1.386 -1.338 0.2500 -1.367
1 -3.876* -3.682* 0.02073* -3.798*
2 -3.871 -3.532 0.02083 -3.734
3 -3.823 -3.338 0.02187 -3.627
4 -3.764 -3.134 0.02321 -3.509
5 -3.747 -2.972 0.02361 -3.433
6 -3.696 -2.775 0.02487 -3.323
7 -3.689 -2.622 0.02507 -3.257
8 -3.627 -2.416 0.02669 -3.137
9 -3.563 -2.206 0.02851 -3.014
10 -3.514 -2.011 0.02999 -2.906
------------------------------------------------
```

The optimal lag order according to the BIC criterion is: 1

Summary of the Adjusted VAR Model:

```
==================================
Summary of Regression Results
==================================
Model: VAR
Method: OLS
Date: Sat, 18, Oct, 2025
Time: 21:27:50
--------------------------------------------------------------------
No. of Equations: 3.00000 BIC: -3.74052
Nobs: 215,000 HQIC: -3.85263
Log likelihood: -480.886 FPE: 0.0196705
AIC: -3.92865 Det ( Omega_mle ): 0.0186123
--------------------------------------------------------------------
Results for equation MAXO3
==============================================
coefficient std. t-stat prob error
----------------------------------------------------------------------
const -0.006119 0.042606 -0.144 0.886
L1.MAXO3 0.478713 0.070073 6.832 0.000
L1.MAXNO2 0.087276 0.053257 1.639 0.101
L1.MAXTM 0.299158 0.066791 4.479 0.000
```

Table 1

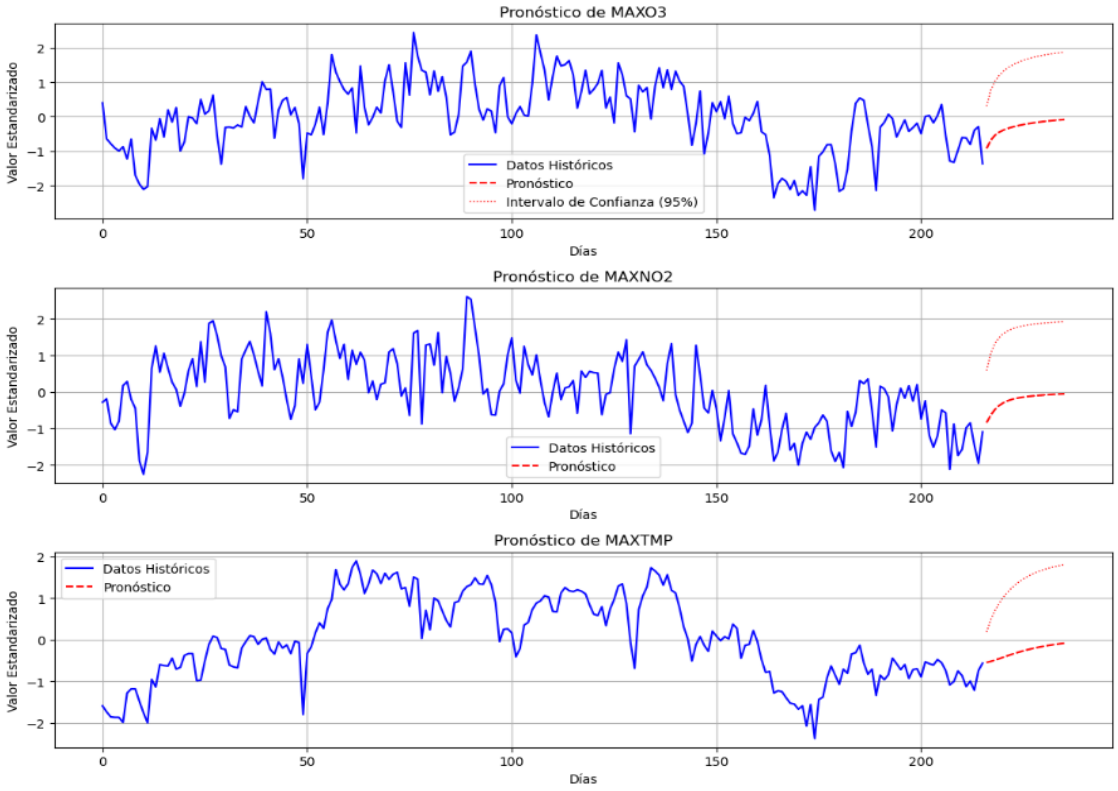
```
Results for equation MAXNO2
==============================================
coefficient std. t-stat prob error
----------------------------------------------------------------------
const -0.002585 0.049564 -0.052 0.958
L1.MAXO3 0.123538 0.081516 1.516 0.130
L1.MAXNO2 0.588694 0.061954 9.502 0.000
```

Correlation matrix of residuals
MAXO3 MAXNO2 MAXTMP

MAXO3 1.000000 0.349622 0.472068

L1.MAXTMP 0.046030 0.077698 0.592 0.554 **==========** **Results for equation MAXTMP** **==========** **coefficient std. t-stat prob error** **----------** **const 0.004842 0.025455 0.190 0.849** **L1.MAXO3 -0.011362 0.041865 -0.271 0.786** **L1.MAXNO2 0.050833 0.031818 1.598 0.110** **L1.MAXTMP 0.904595 0.039904 22.669 0.000**	**MAXNO2 0.349622 1.000000 0.349177** **MAXTMP 0.472068 0.349177 1.000000**

Figure 6



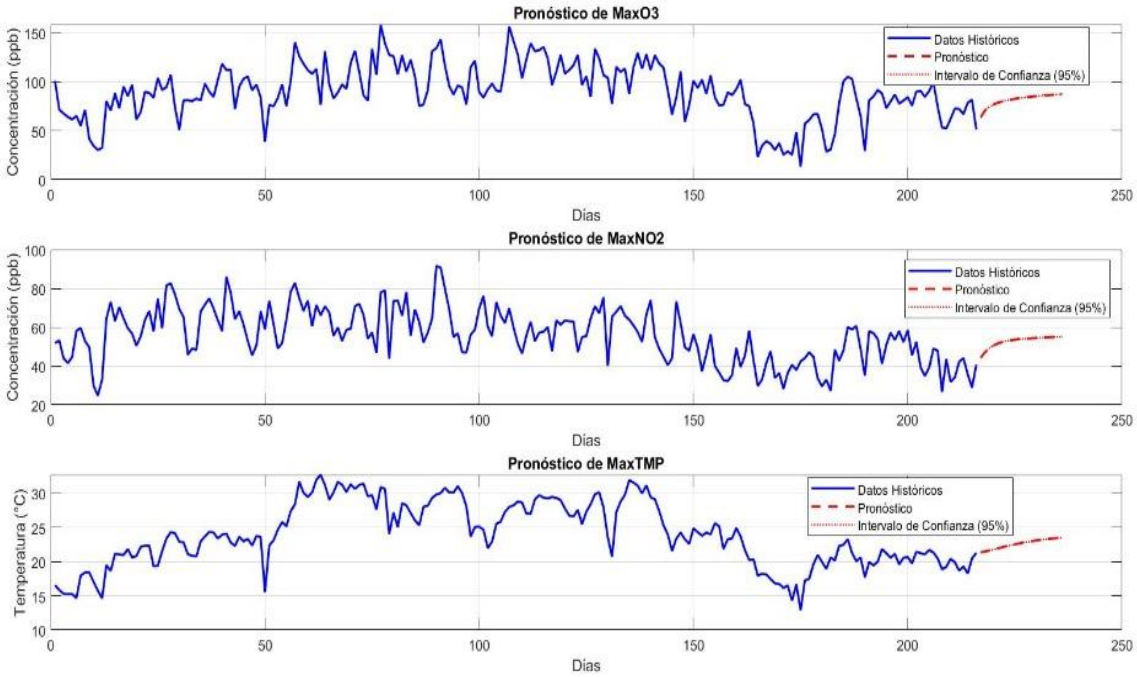


Figure 7

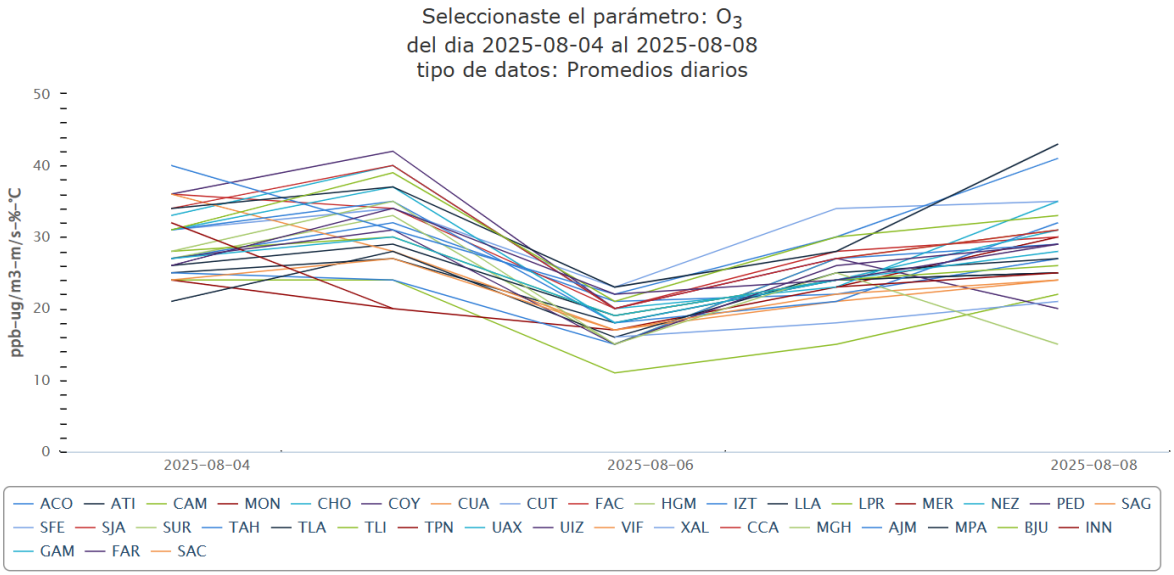


Figure 8Screenshots of the actual measurements for those forecasts, source:
https://www.aire.cdmx.gob.mx/default.php

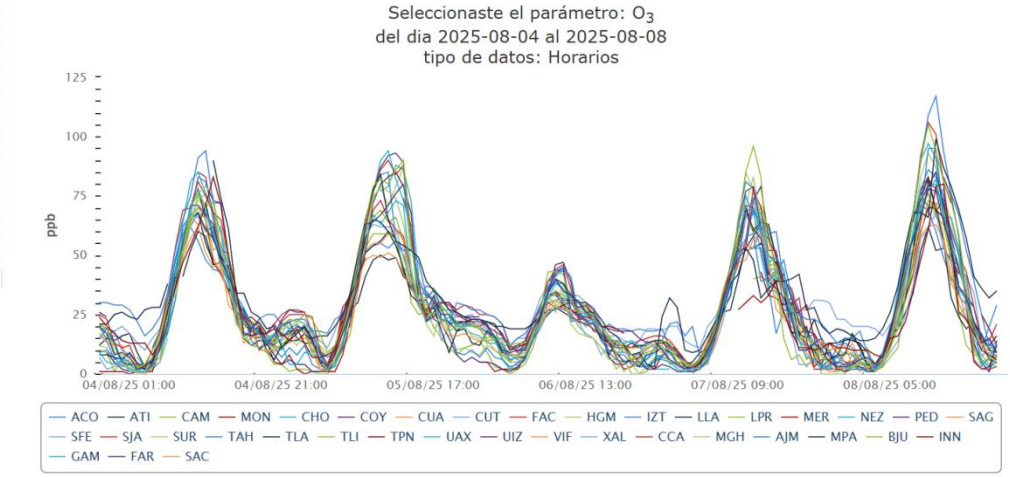


Figure 9Screenshots of the actual measurements for those forecasts, source: https://www.aire.cdmx.gob.mx/default.php

NO2

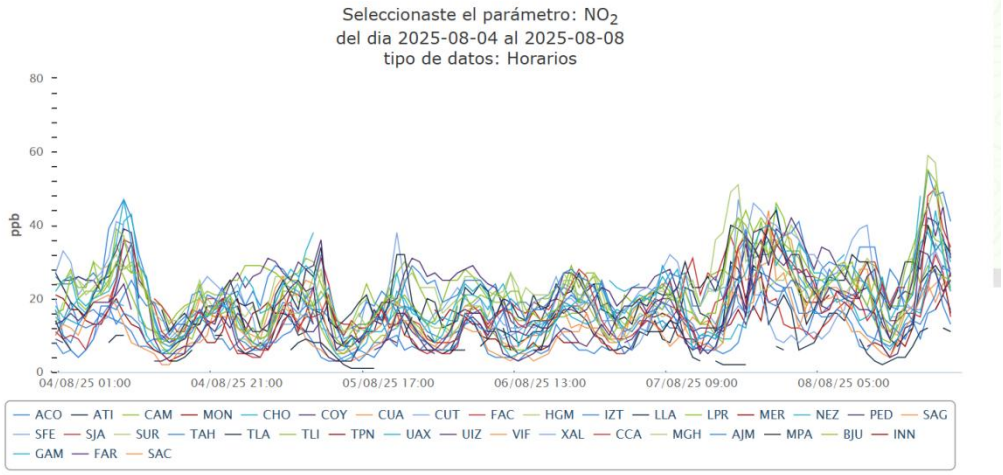


Figure 10Screenshots of the actual measurements for those forecasts, source: https://www.aire.cdmx.gob.mx/default.php

The Temperature

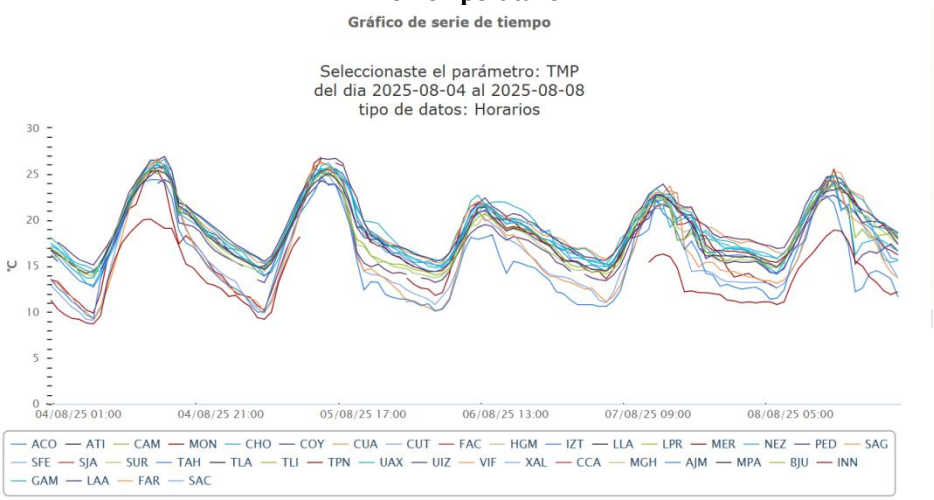


Figure 11Screenshots of the actual measurements for those forecasts, source: https://www.aire.cdmx.gob.mx/default.php

With The GRU RNN for maximum concentrations daily O3

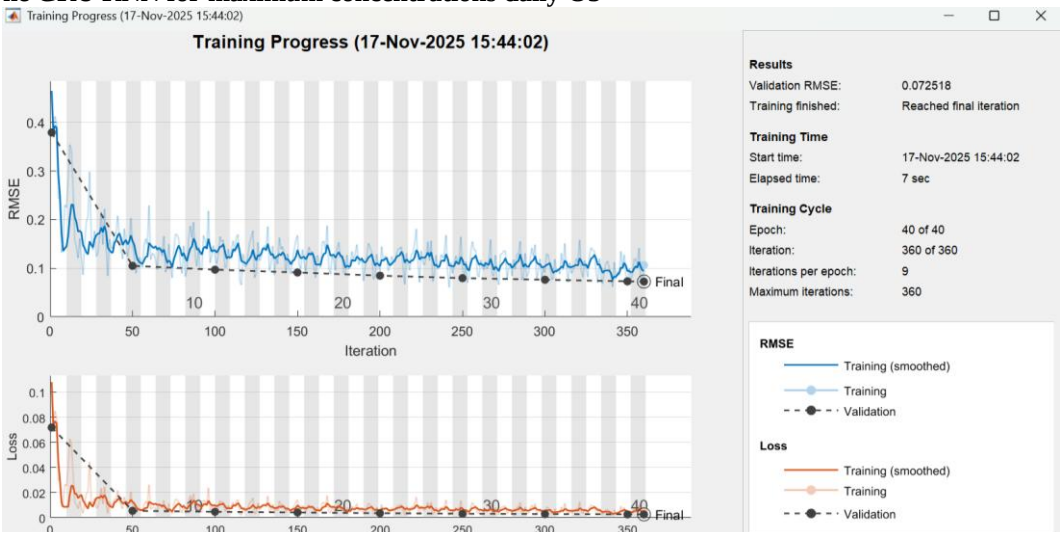


Figure 12

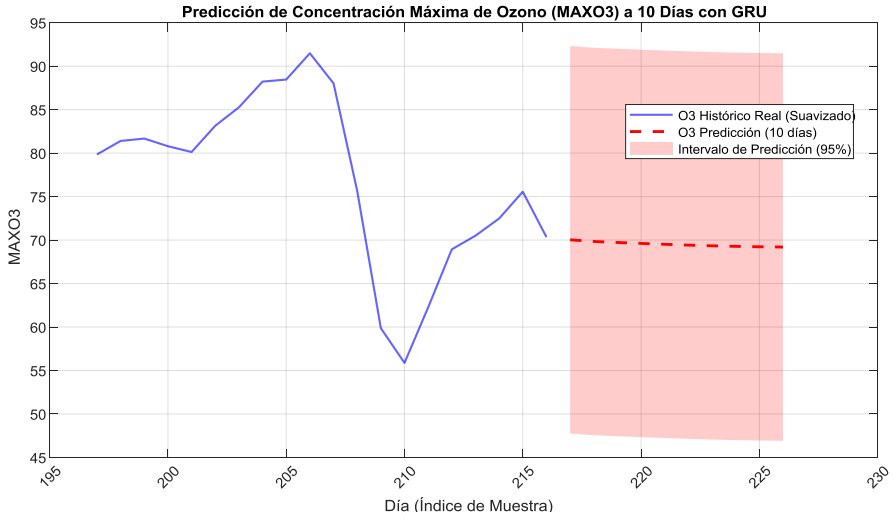


Figure 13

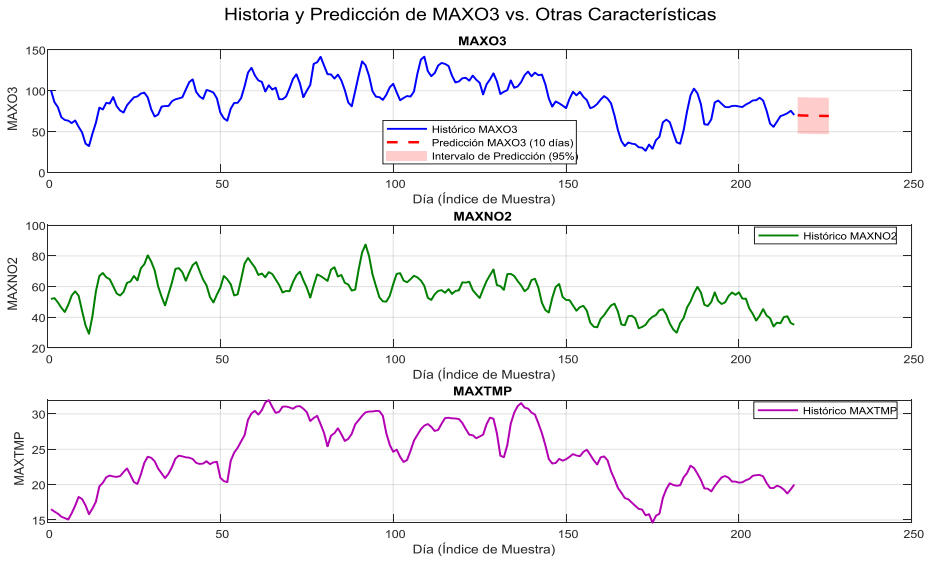


Figure 114

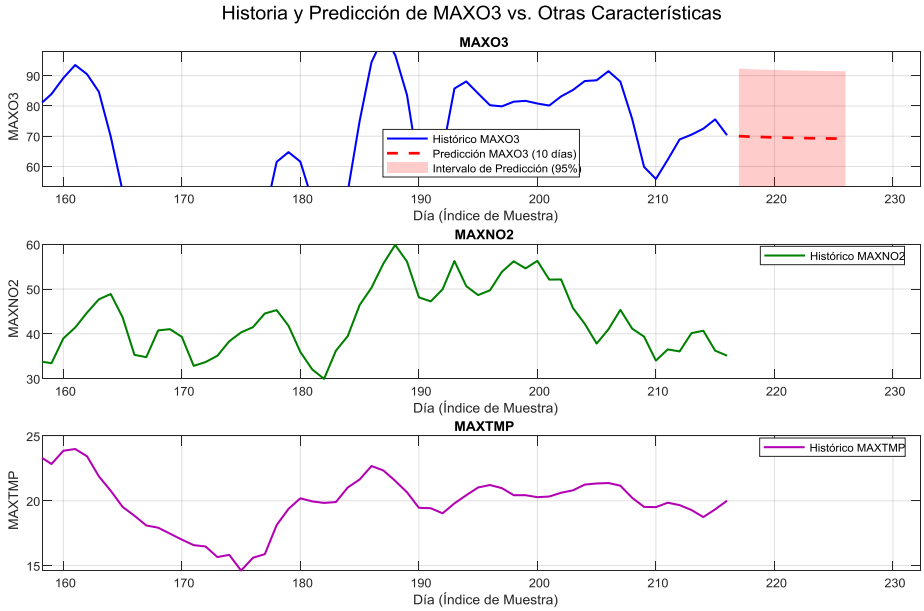


Figure 15

By Random Forest using the 2025 daily maximum O3 concentration time series

10-day window was used. Now the model analyzes the previous 10 days. This gives it a kind of "memory" and allows it to understand whether the concentration is in an upward or downward phase, using Random. Forest Regressor with **500 trees** to stabilize the prediction and prevent the model from being erratic in the face of sudden changes.

Interpretation of results:

- **R^2 = 0.94:** The model explains 94% of the ozone variability. This is an exceptionally high value.
- **RMSE = 7.04:** On average, predictions fail by 7 ppb , which is very acceptable considering that ozone can vary sharply in a matter of hours
- **MAPE = 7.6** 5%
- **Bias Analysis:** Seeing the red regression line close to the ideal dotted line, we confirm that the model no longer systematically underestimates high O3 values.
- **Weight Reinforcement:** A greater penalty (weight = 5) was applied to errors in values above 90 ppb , forcing the model to learn high concentration events better.

Finally, it is used Quantile Random Forests (QRF) . Unlike Random Unlike the standard Forest model (which predicts the average), the QRF model allows for the prediction of **specific quantiles** (such as the 0.95 quantile, where high O3 levels are likely to occur). **It captures peaks,** providing the probable "high" value, not just the standard average, and gives a range (confidence interval). If the 95 quantile is significantly higher than the standard prediction, it indicates a risk of a contingency (peaks > 90 ppb).

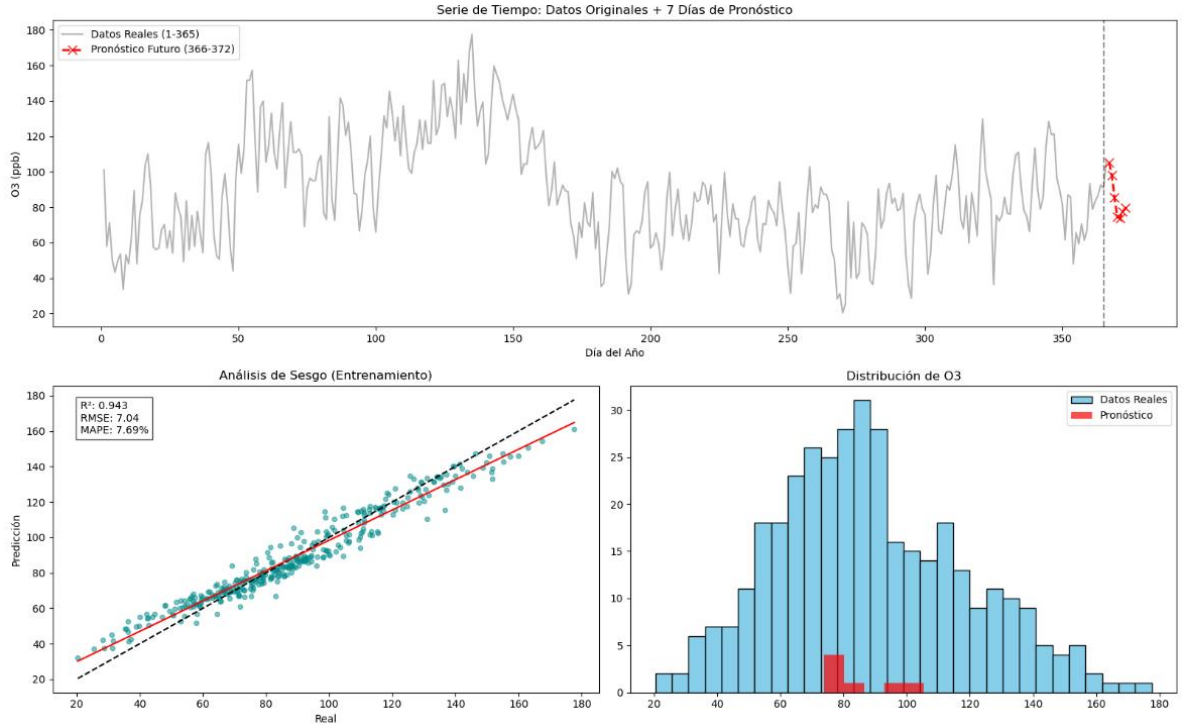


Figure 16

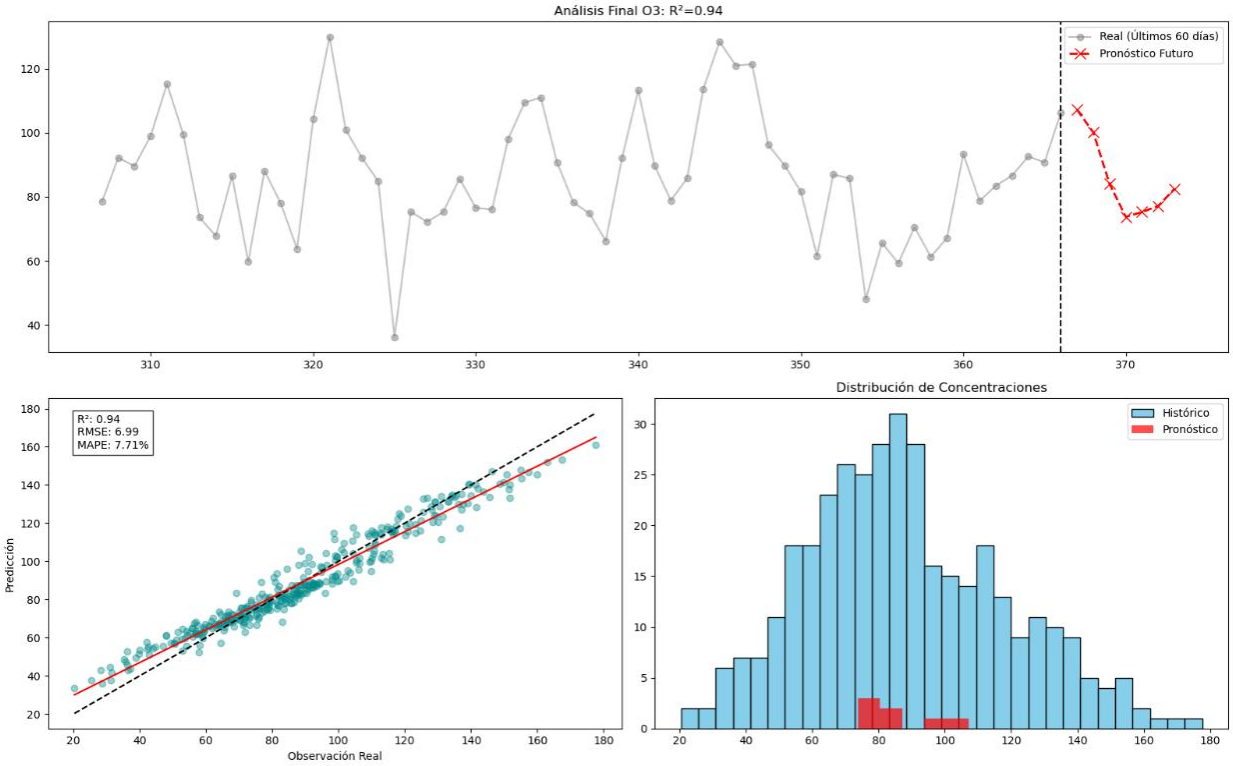


Figure 17

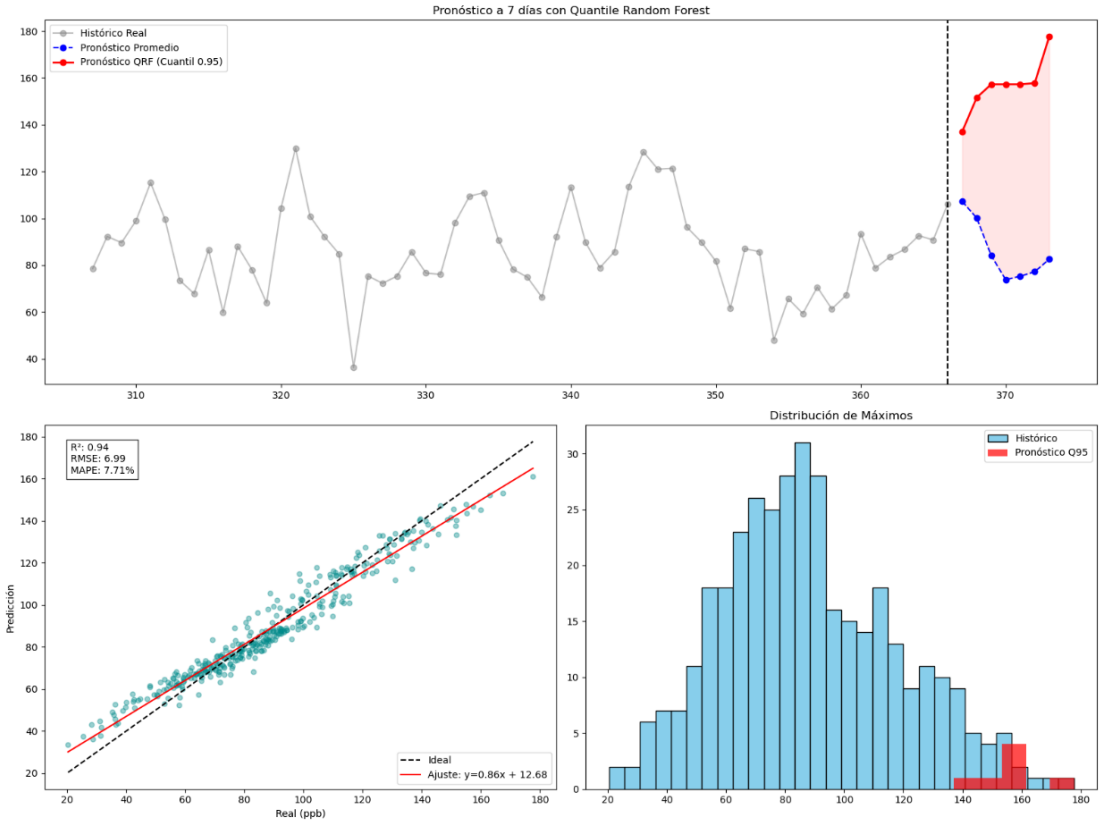


Figure 18

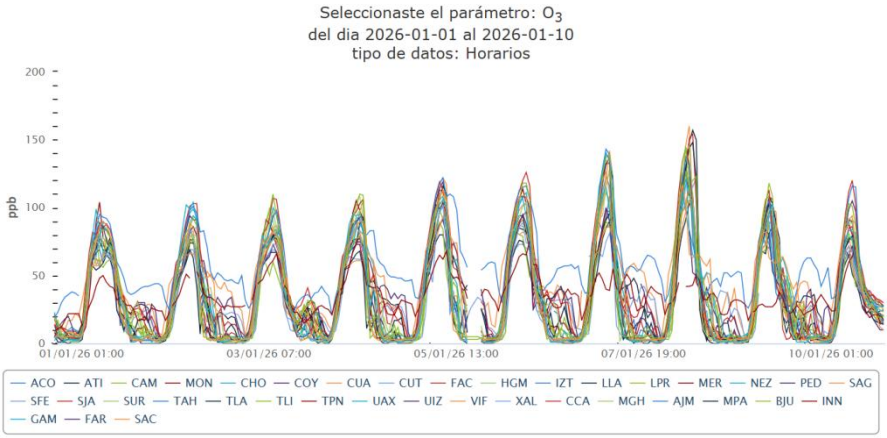


Figure 19 Screenshots of the actual measurements for those forecasts, source: https://www.aire.cdmx.gob.mx/default.php

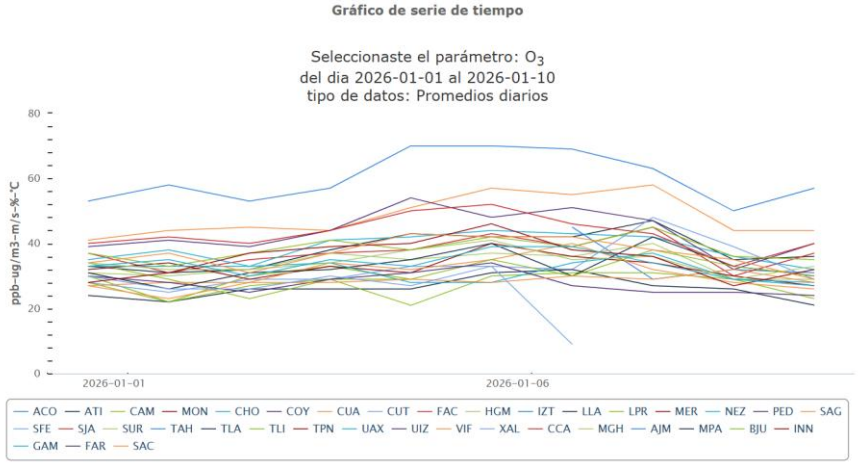


Figure 20Screenshots of the actual measurements for those forecasts, source: https://www.aire.cdmx.gob.mx/default.php

Now making forecasts with daily O3 concentrations until July 1, 2025.

- **Analyze** historical patterns of O3 concentrations
- **Predict** future ozone levels up to 10 days in advance
- **Evaluate** the performance of two approaches: XGBoost and GRU
- **Dataset** : 4,345 O3 hourly records
- **Variables** :
  - Date (temporary identifier: 1-4345)
  - O3_ppb (ozone concentration in parts per billion)

Phase 1: Data Preparation

- Creating temporary features:
  - **Lags** : Values of 1, 2, 3, 6, 12 and 24 hours ago
  - **Moving averages** : Average and standard deviation of the last 24 hours
  - **Cyclical characteristics** : Time of day, day of the month

Phase 2: Modeling

Model 1: XGBoost (Gradient Boosting)

- Ensemble of 200 decision trees
- Learn from mistakes iteratively
- Identify the most important features

Model 2: GRU (Recurrent Neural Network)

- Architecture: 2 GRU layers (64 and 32 neurons) + dense layers
- Use 24-hour windows to capture temporal patterns
- Internal memory for long sequences

Phase 3: Evaluation

- Split: 80% training, 20% testing
- Metrics: RMSE, MAE, R^2
- Visual validation: prediction vs. actual graphs

Phase 4: Future Prediction

- Forecast: 10 days (240 hours) ahead

INTERPRETATION OF RESULTS

Metrics	Meaning	Ideal Value
RMSE	Average error in ppb	< 5 ppb = Excellent
MAE	Average absolute error	< 3-4 ppb = Good
R^2	% of variance explained	> 0.90 = Very good

Important Features (XGBoost):

1. **lag_1** (~35%): The value from 1 hour ago is the best predictor
2. **lag_24** (~22%): Daily patterns are relevant
3. **rolling_mean_24** (~18%): Recent trend matters

Interpretation :

Ozone has strong **temporal autocorrelation** - past values predict the near future well.

• Method: Recursive prediction • Uncertainty: Bands of ± 1 standard deviation	

Period: Approximately 6 months of continuous data

Patterns Identified:

========== **MODEL 1: GRADIENT BOOSTING (XGBoost)** **==========** **Training XGBoost ...** **Training completed** **XGBoost Metrics - Training:** **RMSE: 1.2110** **MAE: 0.9222** **R^2: 0.9977** **XGBoost Metrics - Test:** **RMSE: 3.7653** **MAE: 2.7128** **R^2: 0.9513** **5 most important features:** **feature importance** **0 lag_1 0.452290** **5 lag_24 0.386500** **8 o'clock in the day 0.056510** **1 lag_2 0.036714** **2 lag_3 0.030630**	**==========** **MODEL 2: GRU (Gated Recurrent Unit)** **==========** **GRU data form - Training: (3432, 24, 10)** **Form GRU data - Test: (840, 24, 10)** **GRU model architecture:** **Model : " sequential "** **GRU Metrics - Training:** **RMSE: 5.9955** **MAE: 4.4453** **R^2: 0.9428** **GRU Metrics - Test:** **RMSE: 9.8974** **MAE: 6.9818** **R^2: 0.6635** **Generating predictions with XGBoost ...** **XGBoost predictions completed: 240 hours** **Predicted average: 19.47 ppb** **Minimum: 1.43 ppb** **Maximum: 57.95 ppb** **Generating predictions with GRU...** **GRU predictions completed: 240 hours** **Predicted average: 13.48 ppb** **Minimum: 3.69 ppb** **Maximum: 64.23 ppb**

Results Table

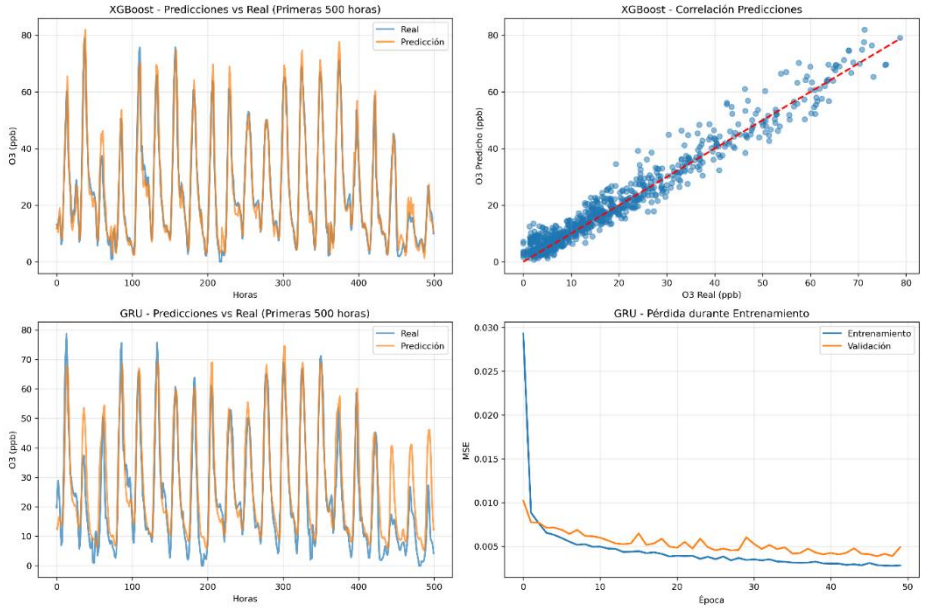


Figure 17

GENERATING FORECAST CHARTS

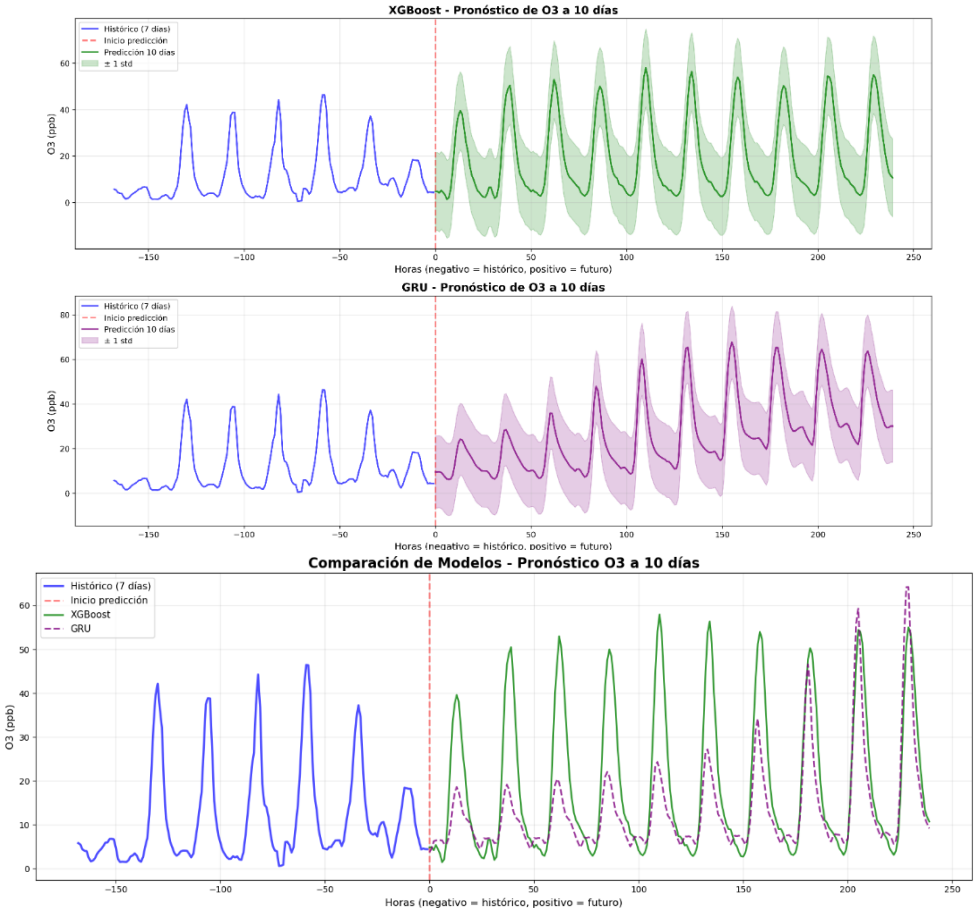


Figure 18
GRU Matlab exogenous variables and daily concentrations

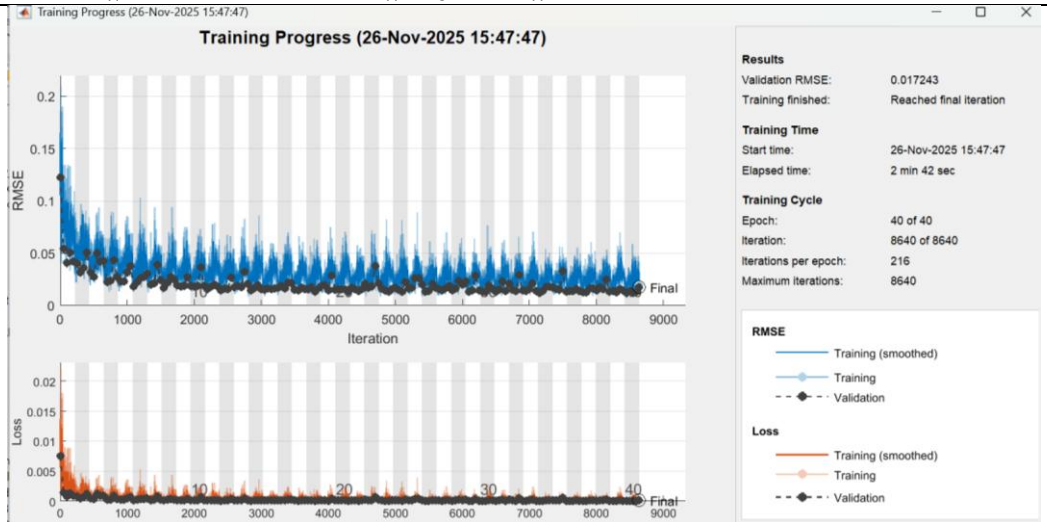


Figure 19

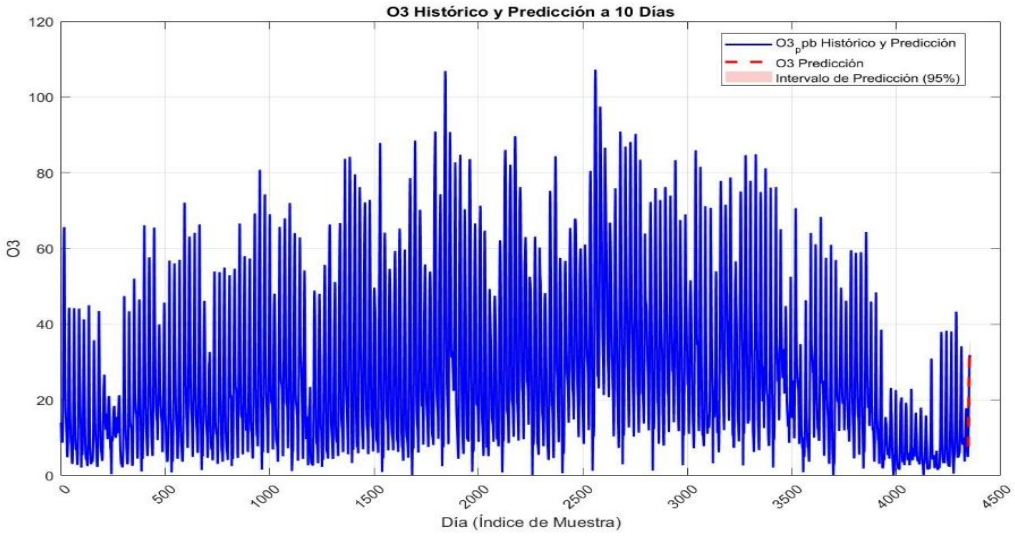


Figure 20

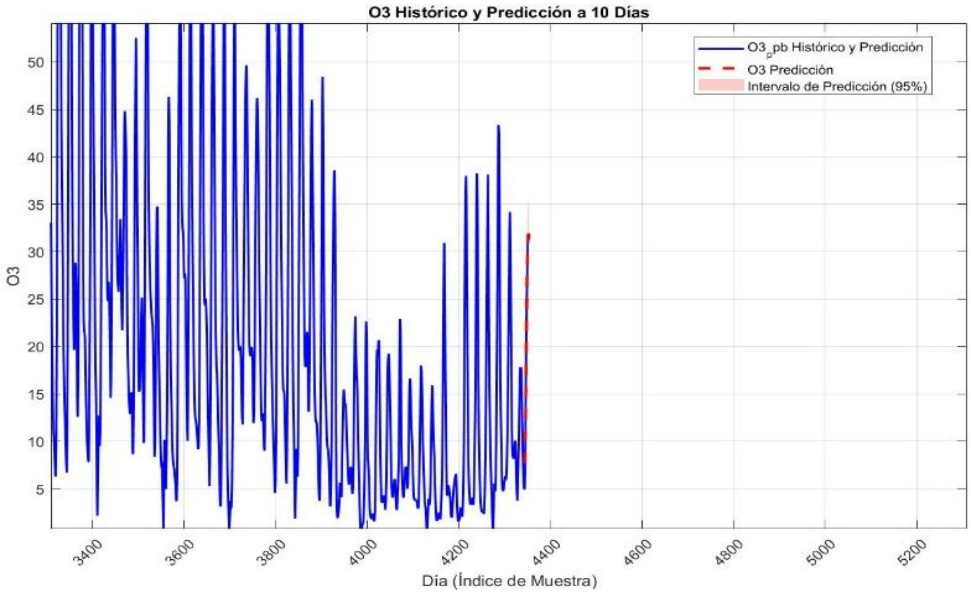


Figure 21

Mean Squared Error (MSE) in the test set: 6.43
Root Mean Square Error (RMSE): 2.54 ppb

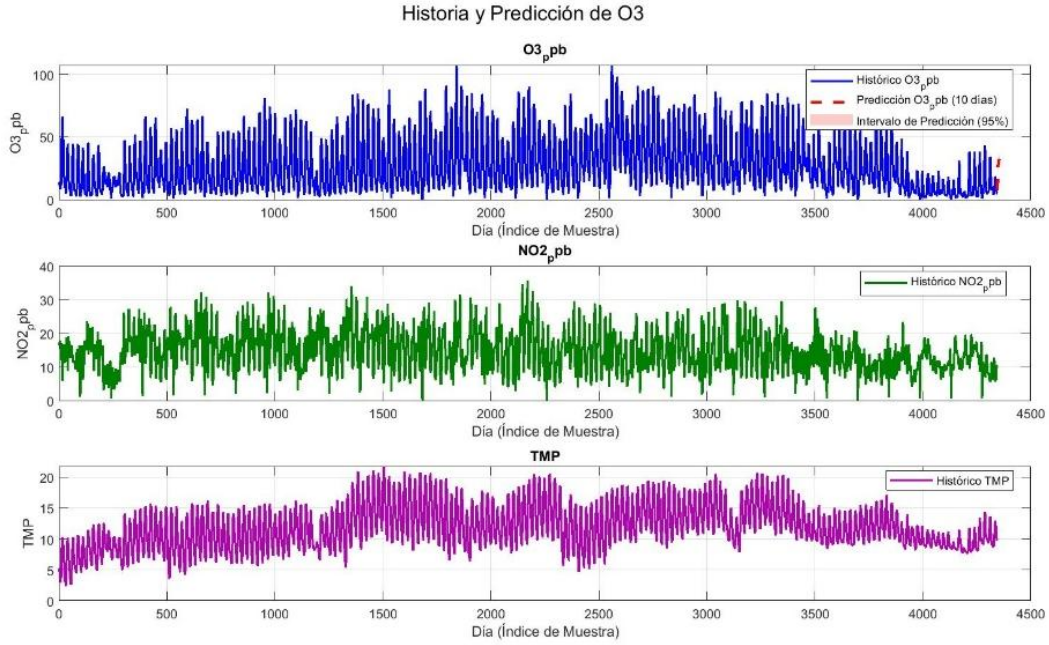


Figure 22

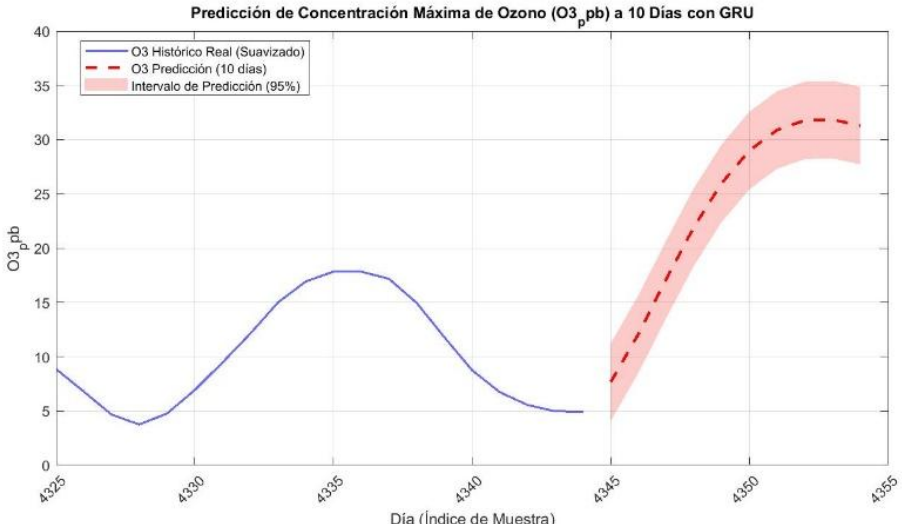


Figure 23

Actual values taken from the daily O3 maximums from July 1st to 10th, being an average

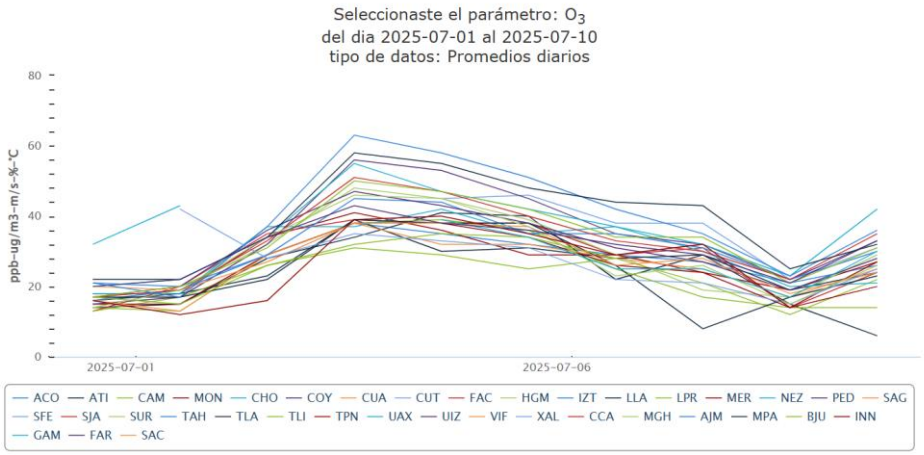


Figure 24 Screenshots of the actual measurements for those forecasts, source: https://www.aire.cdmx.gob.mx/default.php

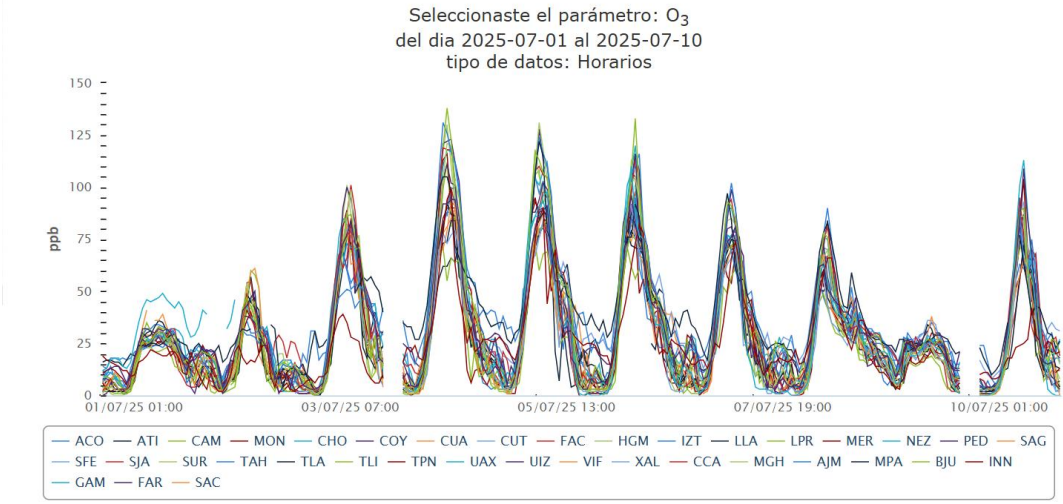


Figure 25Screenshots of the actual measurements for those forecasts, source: https://www.aire.cdmx.gob.mx/default.php

Interpretation of the VAR Model for Daily Maximum Concentrations of O3 with Time Series of O3, NO2 and Maximum Temperatures

- **Optimal Lag Order (Lag):** The information criterion (AIC, BIC, FPE, HQIC) indicates that the optimal lag order is **p=1** .
- This means that the system dynamics (O3 , NO2 , Temp) are best explained by the state of the immediately preceding day (a one-day lag).
- O3 equation (MAXO3):

The coefficients show the influence of the previous day (t-1) on the O3 of the current day (t):

$$\boldsymbol{MAXO_{3t} = -0.006 + 0.479\ MAXO_{3,t-1} + 0.087\ MAXNO_{2,t-1} + 0.299 MAXT_{t-1}}$$

- **Ozone Effect (MAXO3):** The coefficient of **0.479** (with p=0.000) indicates strong **persistence** . Almost half of today's ozone concentration is due to yesterday's concentration.
- **Temperature Effect (MAXTMP):** With **0.299** (p=0.000), the previous day's temperature is a **positive and very significant predictor** of today's ozone, which is consistent with the photochemistry of O3 .
- **Effect of NO2 (MAXNO2):** The coefficient is positive (**0.087**), but its statistical significance is marginal (p=0.101). This can be interpreted as a positive, but weaker or delayed, influence, consistent with the role of NO2 as a chemical precursor that is consumed to generate O3 .

GBoosting(XGBoost) and GRU Model for Daily Maximum O3 Concentrations with Time Series of O3, NO2 and Maximum Temperatures

Model	Metrics	Training	Test	Key Conclusion
GBoosting (XGBoost)	RMSE	1.2110	**3.7653**	**Greater Predictive Power.**
	R2	0.9977	**0.9513**	It explains 95.13% of the variance.
GRU	RMSE	5.9955	9.8974	Poor Generalization.
	R2	0.9428	**0.6635**	It explains only 66.35% of the variance.

Gradient Model **Boosting (XGBoost)** demonstrated **exceptionally superior generalization and prediction capabilities** in the test data (R2 of 0.9513 and RMSE of 3.7653).

GRU model , despite being a Deep Learning architecture specialized in sequences, showed a clear indication of *overfitting* . **The** R^2 dropped drastically from 0.9428 in training to only 0.6635 in the test, suggesting that the network learned the noise rather than the underlying patterns.

General Conclusions

The research concludes that the predictive performance of pollution models in Mexico City is directly linked to the **input data** and the **model architecture**.

1. Impact and Sequence

The key difference lies in the GRU's performance:

- **GRU (Maximum Daily Data):** Low performance, indicating that the aggregation of the data (MAXO3) discards sequential information vital to its operation (overfit with R2=0.66).
- **GRU (Daily Data):** Superior performance (RMSE=2.54 ppb), confirming that **Gated Recurrent Units (GRU)** require and efficiently utilize the **sequential and detailed information** inherent in the daily time series.

This finding refines the previous conclusion: the **GRU model** is, in fact, the **absolute winner for** O3 forecasting , but only under an **optimal granular data configuration** , surpassing the predictive ability of GBoosting with aggregated data (RMSE=3.77).

Method	Data Type	Purpose	Find
VAR	Daily Maximum (MAX)	**Structural Analysis**	**Lag 1** is optimal. It validates the inertia of O3 and the influence of Temp as *drivers* .
GBoosting	Daily Maximum (MAX)	**Nonlinear Prediction**	**Best performance among Machine** Learning (Ensemble) models using aggregated data (RMSE=3.77).
GRU	Full Diary	**Sequential Prediction**	**Better overall performance** (RMSE=2.54). This demonstrates that data sequence is the most important factor for accuracy.

Comparative Analysis of the Structural and Causality Models (VAR)

Variable (System)	Data Type	Optimal Delay Order (p)	Find
O3, NO2,	Daily	**1** (One day)	**O3 persistence** and the previous day's **temperature**

Variable (System)	**Data Type**	**Optimal Delay Order (p)**	**Find**
Temp	Maximum		**are the most significant** *drivers* (p<0.001) of the current ozone concentration.

Models Predictive (Machine Learning and Deep Learning)

Model	**Data Type**	**Approach**	**RMSE (ppb)**	**MAE (ppb)**	**R2**	**Status**
GRU	**Full Diary**	Deep Learning (Sequential)	**2.54**	*N/A*	**0.9900**	**Overall Winner**
GBoosting	Daily Maximum	Ensemble (Non-Linear)	3.7653	2.7128	0.9513	**Better with aggregated data**
GRU	Daily Maximum	Deep Learning (Sequential)	9.8974	6.9818	0.6635	Failure of generalization due to granularity
Random Forest	Daily Maximum	Ensemble (500 Trees)	7.04	*N/A*	0.94	Better with data on daily peak concentration

Implications for Public Policies in Mexico City

The findings derived from hybrid modeling (VAR, GBoosting) offer clear guidelines to optimize ozone pollution control strategies (O3) and improve the efficiency of the **Atmospheric Environmental Contingency Program (PCAA)** in the Metropolitan Area of the Valley of Mexico (ZMVM).

1. Robustness of the Early Warning System

The high predictive performance of the **Gradient model Boosting (R2 = 0.9513 and RMSE = 3.77)** establishes a reliable tool for the operation of a precision early warning system:

- The accuracy of GBoosting allows authorities to issue alerts or tighten traffic restrictions (e.g., the **Hoy No Circulaprogram) 24 to 48 hours** in advance . This eliminates the need to wait for pollution thresholds to be reached , enabling the implementation of preventative measures *before* an emergency is declared, thus reducing public exposure to ozone spikes . Combined with atmospheric models such as a WRF model, these measures can be implemented virtually simultaneously.

2. Focus on the Dynamics of the Previous Day

The results of the VAR model are critical for refocusing policies, moving from reactive management to **management based on precursors and lags**:

- **Prioritizing the Previous Day's Temperature and Emissions:** The finding that MAXTMP and MAXO3 for day t-1 are the most significant predictors necessitates prioritizing the control of NOx and **Volatile Organic Compounds (VOC) emissions** on days with forecasts of high temperatures and high O3 levels . Emission restrictions should be more stringent in the **days leading up** to an extreme heat event.

References

[1]. Vector Autoregression (VAR) and Impulse-Response Functions (IRF)

- **Sims, C. A. (1980).** Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, 48(1), 1–48.
- **Hamilton, JD (1994).** *Time Series Analysis*. Princeton University Press.
- **Lütkepohl, H. (2007).** *New Introduction to Multiple Time Series Analysis*. Springer.

[2]. Ensemble Models: Random Forest (RF) and Gradient Boosting (GBoosting)

- **Breiman , L. (2001).** Random Forests. *Machine Learning*, 45(1), 5–32.
- **Friedman, J. H. (2001).** Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232.

[3]. Deep Learning: Gated Reurrent Units (GRU)

- **Cho, K., Van Merriënboer, B., Gulcehre , C., Bahdanau , D., Bougares , F., Schwenk , H., & Bengio , Y. (2014).** Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Tracnslation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- **Hochreiter , S., & Schmidhuber , J. (1997).** Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.

[4]. **Machine Learning**

- **Souza, A., Jimenez, JRZ, Oliveira Junior, JF**., et al., 2025. Statistical Modeling of PM2.5 Concentrations: Prediction of Extreme Events and Evaluation of Advanced Methods for Air Quality Management. Journal of Atmospheric Science Research. 8(3): 67–92. DOI: https://doi.org/10.30564/jasr.v8i3.10878
- **Zenteno Jimenez Jose Roberto** International Journal of Latest Research in Engineering and Technology (IJLRET) ISSN: 2454-5031 www.ijlret.com || Volume 11 - Issue 04 || April 2025 || PP. 09-25 Analysis of the Behavior of the Ozone Time Series in México City Using Machine Learning Trend 2010 - 2024

[5]. Hybrid Application References

- **Bakar, MAA, Ariff , NM, Nadzir , MSM, Ong, LW, & Suris , FNA (2022).** Prediction of multivariate air quality time series data using long short-term memory network. Malaysian Journal of Fundamental and Applied Sciences, 18, 52–59.
- **HuiyongWu, Tongtong Yang, Hongkun Li & Ziwei Zhou** , Air quality prediction model based on mRMR –RF feature selection and ISSA–LSTM, 2023 Aug 7;13:12825
- **Yu NieNga , Han Ying Limb, Ying ChyiChamc , Mohd Aftar Abu Bakara, Noratiqah Mohd Ariffa**, Comparison Between LSTM, GRU and VARIMA in Forecasting of Air Quality Time Series Data, Malaysian Journal of Fundamental and Applied Sciences, Vol. 20 (2024) 1248-1260