

Article

Not peer-reviewed version

Prediction of Air Pollutants Based on Time-Weighted Ensemble Model and Adaptive Air Quality Index Model

[Borui Wang](#) , Chenjie Gong , [Chao Liu](#) ^{*} , Jiahe Yang , Huili Huang

Posted Date: 16 January 2026

doi: 10.20944/preprints202601.1239.v1

Keywords: time-weighted ensemble model; adaptive air quality index; KNN interpolation; unit transformation; normalization; mutual information; Spearman's rank correlation; Kendall's tau correlation



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Prediction of Air Pollutants Based on Time-Weighted Ensemble Model and Adaptive Air Quality Index Model

Borui Wang¹, Chenjie Gong², Chao Liu^{2,*}, Jiahe Yang³ and Huili Huang⁴

- ¹ Mathematics and Statistics College,Northeastern University of Qinhuangdao
- ² SINOPEC (HONG KONG) LIMITED
- ³ Sydney Smart Technology College, Northeastern University at Qinhuangdao
- ⁴ China Resources (Holdings) Co., Ltd.
- * Correspondence: liuchao@mail.neu.edu.cn; Tel.: +86-0335-8056200

Abstract

Air quality makes a huge difference in human health, ecological environment, economic development, and global climate governance. This study introduced **Time-Weighted Ensemble model** into the air quality prediction model and achieved good results. The prediction results were consistent with reality, and the R-square of prediction is 0.54, providing a new reference for people to avoid air pollution. And because of the the original Air Quality Index (AQI) has limited using scope and results are inaccurate, this thesis establish a brandnew evaluation system, called **Adaptive Air Quality Index (AAQI)**, which takes concentration, correlation, time, and cooperation into consideration. It is more comprehensive and advanced than the existing system. Data on six pollutants were collected from six cities, namely Brasilia, Cairo, Dubai, London, New York, and Sydney, and then preprocessed the above data using **KNN interpolation, Unit transformation and normalization**, and calculated the correlations among them by using **Mutual Information, Spearman's Rank Correlation** and **Kendall's Tau Correlation**. Afterwards, we incorporated it into the AAQI and obtained their air quality. Among them, Sydney had the best air quality, while Dubai and Cairo had relatively poor air quality. This research should be promoted and applied in air quality monitoring in real life.

Keywords: time-weighted ensemble model; adaptive air quality index; KNN interpolation; unit transformation; normalization; mutual information; Spearman's rank correlation; Kendall's tau correlation

1. Introduction

Air pollution is one of the biggest environmental risks affecting health. 99% of the global population resides in areas that do not meet the World Health Organization's air quality standards.[1] If people are exposed to air pollution for a long time, it can cause respiratory diseases, cardiovascular diseases, neurological damage, and ultimately shorten their lifespan.[2] Good air quality is the foundation of human health and ecological balance. Therefore, predicting air quality accurately in order to help people avoiding air pollution is necessary and significant. After comparing various machine learning and deep learning models, we found that no model has shown a significant advantage in prediction. Consequently, this article based on these models create an time related ensemble model to predict the concentration of air pollutants, and calculates the AAQI value based on the prediction results. Through this new model, this study not only provides new ideas for air quality monitoring, but also provides practical tools for the establishment of public health warning systems.

Traditional air quality monitoring typically relies on the Air Quality Index (AQI). However, existing AQI models have many limitations. It was found that the original Air Quality Index (AQI) only took the maximum sub-index and did not consider the health effects of multiple pollutants,

neglecting the synergistic effect of different pollutants. Moreover, when the pollutant concentration exceeded the maximum limit of 500, AQI could not further reflect the degree of pollution and only displayed "explosive" information, resulting in information distortion and significant limitations. Therefore, these issues make it difficult for AQI to comprehensively and accurately reflect the impact of air quality on humane health and it is necessary to create a new model to measure air quality. Air Quality Index formula:

The sub-index for each pollutant is calculated based on its concentration and predefined break-points. The formula for the sub-index of a given pollutant is:

$$I_p = \frac{C_p - C_{p,\min}}{C_{p,\max} - C_{p,\min}} \times (I_{p,\max} - I_{p,\min}) + I_{p,\min}$$

Where: I_p is the sub-index for pollutant p , C_p is the concentration of pollutant p (usually in micrograms per cubic meter, $\mu\text{g}/\text{m}^3$), $C_{p,\min}$ and $C_{p,\max}$ are the minimum and maximum concentrations for pollutant p , typically based on the regulatory standard limits, $I_{p,\min}$ and $I_{p,\max}$ are the minimum and maximum sub-index values for pollutant p , which are usually between 0 and 500.[3]

The overall AQI is determined by selecting the maximum sub-index among all pollutants. If there are multiple pollutants (e.g., PM2.5, PM10, CO, NO2, SO2, O3), the AQI is calculated as:

$$\text{AQI} = \max(I_{\text{PM2.5}}, I_{\text{PM10}}, I_{\text{CO}}, I_{\text{NO2}}, I_{\text{SO2}}, I_{\text{O3}})$$

A critical evaluation of air quality index models (1960–2021) analyses the strengths and flaws of all the AQI models developed so far, it is recommended to develop a more reliable, extensible, and comparable AQI model to be employed as an executive tool for designing strategic pollution abatement programs to preserve public health.[4] To address these issues, this study proposes an innovative air quality assessment model - Adaptive Air Quality Index (AAQI). The AAQI model combines pollutant concentration, correlation between pollutants, time factors, and synergistic effects of pollutants on the basis of the original AQI. This model can provide a more comprehensive and accurate assessment of air quality than traditional AQI, especially in cases where pollutant concentrations exceed the standard. AAQI can better reflect changes in air quality.

This study is based on Kaggle data, which includes air quality data for six cities (Brasilia, Cairo, Dubai, London, New York, Sydney), including segmented hourly data for six indicators of CO, NO2, SO2, O3, PM2.5, PM10, and AQI for the whole year of 2023 and 2024, totaling 105200 data pieces.[5,6] A new air quality index model, AAQI (Adaptive Air Quality Index), was established to comprehensively measure the impact of air quality on human health by studying the effects of pollutants on health, adjusting weights for correlation, time weights, and synergistic effects.

2. Data Processing

2.1. KNN Interpolation for Missing Values

Before conducting data analysis, this article cleans and preprocesses the raw data to ensure its quality and reliability. The first step is to handle missing values. Due to the strong correlation between air quality data and time, the concentration of pollutants fluctuate dramatically at different time. Traditional mean imputation methods (such as filling missing values with global means) cannot effectively capture the temporal characteristics of the data. Consequently the KNN interpolation methods is most suitable for simulating missing data using surrounding data.

KNN interpolation (K-Nearest Neighbors Interpolation) is a non parametric method based on neighboring samples. The core idea is to calculate the distance between the interpolation point and the known data point, find the nearest K neighbors, and use the values of these neighbors for weighted or average estimation. If we have a missing value $\hat{X}_i$ for feature X_i we use the following formula to impute it based on the k nearest neighbors.

$$\hat{X}_i = \frac{1}{k} \sum_{j \in N_k(i)} X_j \quad (1)$$

where $N_k(i)$ is the set of the k nearest neighbors to sample ($k = 5$), X_j is the feature value of the j -th neighbor, k is the number of nearest neighbors, $\hat{X}_i$ is the imputed value based on the values of the neighbors.[7]

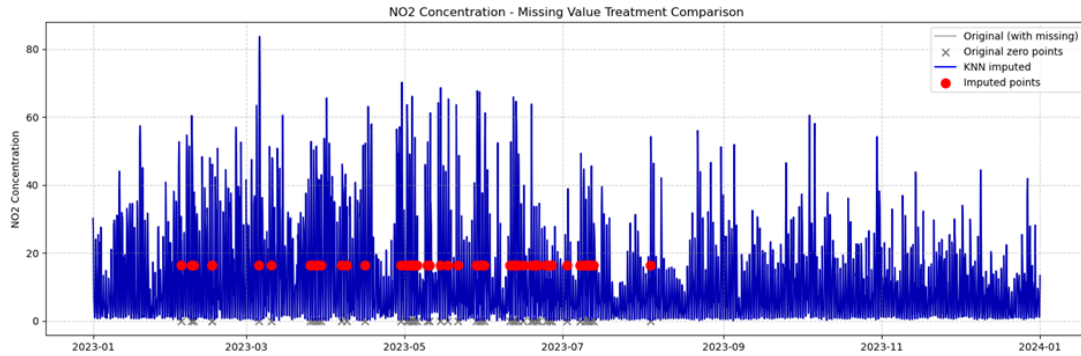


Figure 1. Comparison of KNN interpolation effects.

2.2. Unit Transformation

In the dataset, the concentration of CO is in ppm and the other pollutants are in $\mu\text{g}/\text{m}^3$. If the raw values with different units are directly substituted into the normalization process, the significant difference in magnitudes will cause the model to overly favor the large values of CO, ignoring the key information of PM. This not only affects the prediction accuracy but also undermines the fairness of the AAQI (Air Quality Index) weighting calculation.

The general conversion formula is (Using Brasilia as an example):

$$\text{Concentration}(\mu\text{g}/\text{m}^3) = \text{Concentration}(\text{ppm}) \times \frac{\text{Molecular Weight (g/mol)}}{\text{Gas Constant (L} \cdot \text{atm/mol} \cdot \text{K)} \times \text{Temperature (K)}} \times \frac{P}{P_0} \times 10^6$$

Where: P is the atmospheric pressure in Brasilia (95 kPa = 0.95 atm), P_0 is the standard atmospheric pressure (101.325 kPa = 1 atm), T is the temperature (295 K).

$$\text{Concentration}(\mu\text{g}/\text{m}^3) \approx \text{Concentration}(\text{ppm}) \times 1.148$$

Therefore, at the atmospheric pressure and temperature in Brasilia, 1 ppm of CO is approximately $1.148 \mu\text{g}/\text{m}^3$. [8]

2.3. Data Normalization

Data normalization is aiming to eliminate the influence of different index sizes, the min-max normalization method is used to normalize all data to the [0,1] interval. The standardized formula is as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (2)$$

Where X' is the normalized value, $X_{\min}$ is the minimum value of features, $X_{\max}$ is the maximum value of features, X is the Original eigenvalue. [9]

2.4. Calculation of Nonlinear Correlations

After that, researchers want to understand the interactions and impact relationships between various pollutants, so we calculate nonlinear correlations. Because traditional correlations calculation method such as Pearson correlation applies to the linear relationships, three methods are employed to compute the correlation between pollutants in this research, which is beneficial to measure nonlinear relationships.

Mutual Information measures the amount of information shared between two variables, capturing both linear and nonlinear dependencies. MI is particularly useful for identifying relationships that

might be overlooked by traditional correlation methods. The MI between pollutants X and Y is computed using the following formula:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Where: $p(x,y)$ is the joint probability distribution of the two variables, and $p(x)$ and $p(y)$ are the marginal distributions.[10]

Spearman's correlation assesses the monotonic relationship between two variables by comparing their ranks rather than their actual values. It is particularly useful for capturing nonlinear, monotonic dependencies. The Spearman correlation coefficient ρ is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where: d_i is the difference in ranks for each data pair, and n is the number of data points.[11]

Kendall's Tau measures the degree of concordance and discordance between pairs of data points. It is also based on ranks, but unlike Spearman's, it uses a pairwise comparison approach. The Kendall Tau coefficient τ is computed as:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$$

where C is the number of concordant pairs and D is the number of discordant pairs. [12]

Combined Correlation Matrix To capture the full range of dependencies between pollutants, a weighted combination of the three correlation methods is used. The final combined correlation matrix M_{combined} is calculated as:

$$M_{\text{combined}} = 0.6 \times M_{\text{MI}} + 0.3 \times M_{\text{Spearman}} + 0.1 \times M_{\text{Kendall}}$$

This combined approach allows for a more nuanced understanding of the interdependencies between pollutants, integrating both linear and nonlinear relationships.

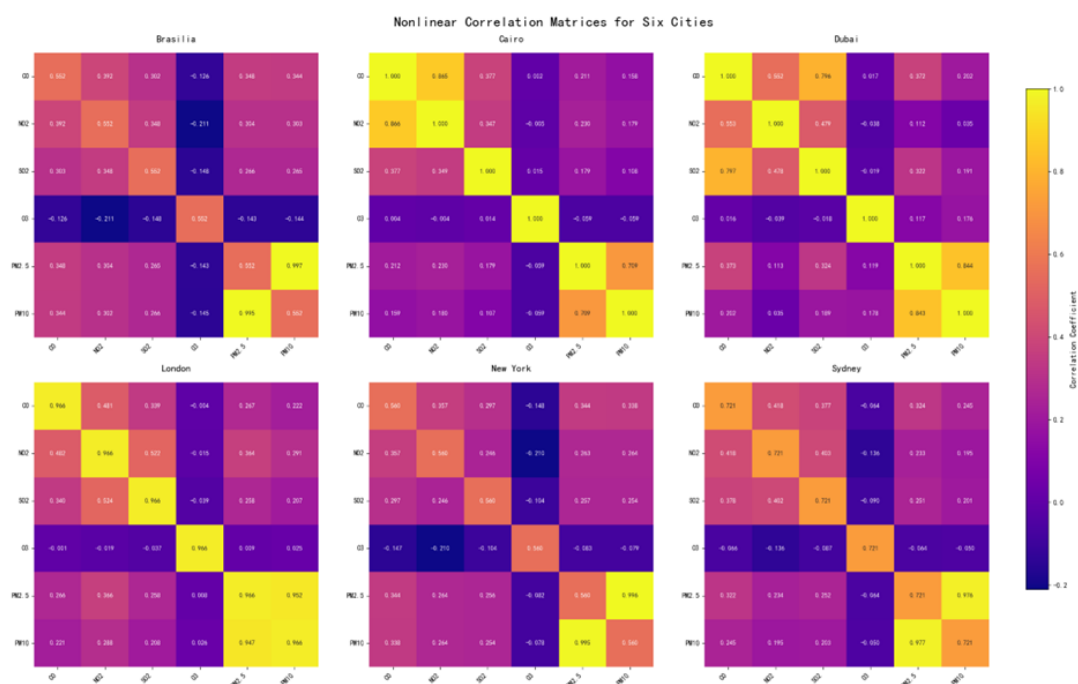


Figure 2. Heat map of air pollutant correlation coefficient from six different cities.

3. Predicting Air Quality by Time-Weighted Ensemble Model

In this section, we take LSTM, DNN, RF and ARIMA models into considerations, which are highly sensitive to time and have shown good results in series predictions before.

3.1. Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) designed to effectively capture long-term dependencies in sequential data. It can automatically learn the optimal temporal lag structure from the data, eliminating the need for manual lag selection and model complex temporal and nonlinear relationships, by incorporating nonlinear activation functions, which is very suitable for predict the air quality. An LSTM cell contains three key gates — the input gate, forget gate, and output gate — which regulate the flow of information. The gating structure allows LSTM to selectively retain or forget information, enhancing robustness when dealing with noisy or irregular input data. The model formulas are as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \odot \tanh(C_t) \end{aligned}$$

Where, x_t is the input at time step t , h_t is the hidden state, C_t is the cell state, f_t , i_t , and o_t denote the forget, input, and output gates respectively, σ is the sigmoid activation function, and $\odot$ represents element-wise multiplication.[13]

3.2. Deep Neural Networks (DNN)

Deep Neural Networks (DNN) is a type of artificial neural network with multiple layers between the input and output layers. DNN is good at capturing complex, non-linear relationships in data. Air quality data often exhibits non-linear trends, seasonal variations, and other complex behaviors. DNNs can model these non-linear relationships without explicitly needing to define them, which can lead to better predictive performance. The model formulas are as follows:

$$\begin{aligned} a^{(l)} &= \sigma(W^{(l)} \cdot a^{(l-1)} + b^{(l)}) \\ \text{MSE} &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ W^{(l)} &= W^{(l)} - \eta \cdot \frac{\partial L}{\partial W^{(l)}} \\ b^{(l)} &= b^{(l)} - \eta \cdot \frac{\partial L}{\partial b^{(l)}} \\ \hat{y} &= W^{(L)} \cdot a^{(L-1)} + b^{(L)} \end{aligned}$$

Where, $W^{(l)}$ is the weight matrix for layer l , $a^{(l-1)}$ is the output of the previous layer (for the first layer, it is the input X), $b^{(l)}$ is the bias vector for layer l , σ is the activation function ($\sigma(x) = \max(0, x)$) (ReLU) for hidden layers, linear for the output layer), $a^{(l)}$ is the output (activation) of layer l , y_i is the true value (target), $\hat{y}_i$ is the predicted value from the model, n is the number of data points, η is the learning rate, $\frac{\partial L}{\partial W^{(l)}}$ and $\frac{\partial L}{\partial b^{(l)}}$ are the gradients of the loss function with respect to the weights and

biases, respectively, L is the final (output) layer of the network, $a^{(L-1)}$ is the output of the last hidden layer, $W^{(L)}$ and $b^{(L)}$ are the weight matrix and bias vector of the output layer.[14]

3.3. Random Forest (RF)

Random Forest is an ensemble learning method based on multiple decision trees, which can capture complex non-linear relationships and interactions between features. It is less sensitive to overfitting and works well with noisy data, making it reliable for real-world air quality predictions where noise and outliers are common. The random forest regression formula is:

$$\hat{y}_i = f_i(X)$$

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N \hat{y}_i$$

Where, $\hat{y}_i$ is the prediction from the i -th decision tree, X is the input feature vector, f_i is the function learned by the i -th decision tree, which outputs a predicted value based on, $\hat{y}$ is the final prediction, N is the number of trees in the forest.[15]

3.4. Autoregressive Integrated Moving Average (ARIMA)

Autoregressive Integrated Moving Average Model (ARIMA) is a statistical model used for time series prediction. By combining autoregression (AR), differencing (I), and moving average (MA), ARIMA is able to capture trends and seasonal variations in data. The preponderance of ARIMA is strong adaptability, which can handle various time series data like . The general ARIMA(p, d, q) model can be written as:

$$y_t = c + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \epsilon_t$$

Where, y_t is the actual value of the time series at time t , c is a constant (optional, sometimes set to 0), ϕ_i are the parameters of the autoregressive part (AR), ϵ_t is the error term at time t , also known as the white noise (residuals), θ_j are the parameters of the moving average part (MA), ϵ_{t-j} are the lagged residuals (errors) from previous time steps, p is the number of lag terms in the autoregressive model, q is the number of lag terms in the moving average model.[16]

The dataset of above four models is divided according to the following method. The training set is from January to October of 2023 (accounting for 83.2%), which is used to learn the seasonal and daily variation patterns of pollutants. Validation set is November of 2023 (accounting for 8.2%), using to optimize hyperparameters of LSTM, DNN, and random forest (such as the number of hidden layer nodes, learning rate, dropout rates, batch size, the number of estimators, max depth and so on). Test set is December of 2023 (accounting for 8.4%), verifying the generalization ability of the model on recently unknown data.

Because different models have different outcomes, some focus on MAE and some focus on R-squared. In order to improve the accuracy of prediction results, we have created a ensemble model with time weights. We have also invented a new formula for calculating model performance, which adjusts the time weights of each model by maximizing model performance parameters. The performance score combines MAE and R^2 :

$$\text{Performance Score} = \frac{1}{1 + \text{MAE}} \times \frac{R^2 + 1}{2}$$

Where, MAE is mean absolute error, which measures accuracy. R^2 measures goodness of fit.

3.5. Time-Weighted Ensemble Model

The Time-Weighted Ensemble combines the predictions from LSTM, DNN, and Random Forest models based on hourly optimized weights.

Considering that the effect of direct weighting is not much better than a single model, we adopt the hourly weighting method to improve the performance of the ensemble model. The weights for each model are optimized for each hour to minimize the MAE for the ensemble prediction:

$$\hat{y}_{ensemble} = \sum_{i=1}^n w_i \cdot \hat{y}_{i,h}$$

Where w_i are the weights, and $\hat{y}_{i,h}$ are the predictions from each model.

In addition, if one model performs significantly better than others (based on the performance ratio), that model is used exclusively for that hour.

4. Predicting Result of Multiple Model

Using Time-Weighted Ensemble model, this study predicted the changes in NO2 concentration throughout 2024 and the NO2 concentration at various times on January 1, 2024, by using the data from December 29, 30 and 31, 2023. The calculated results and actual pollutant concentrations are as follows:

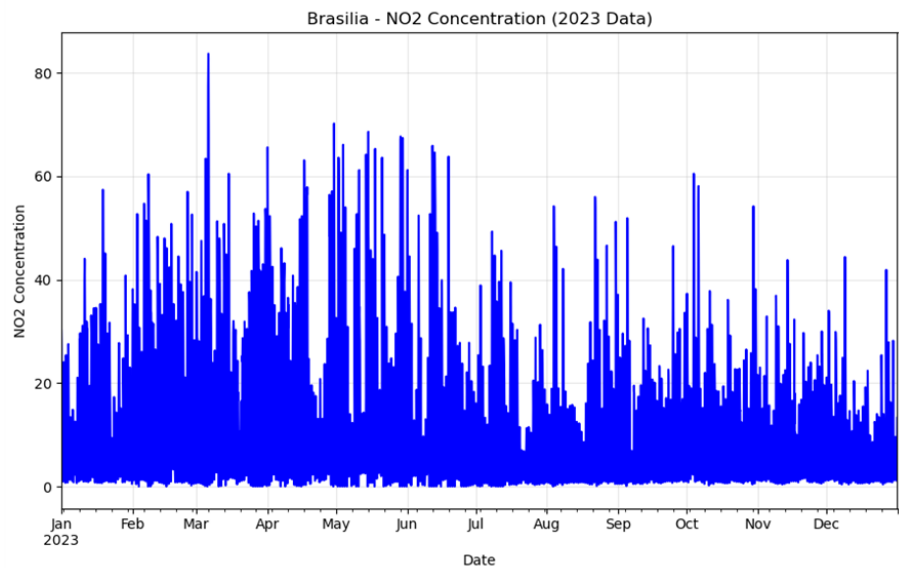


Figure 3. CO concentration throughout 2023.

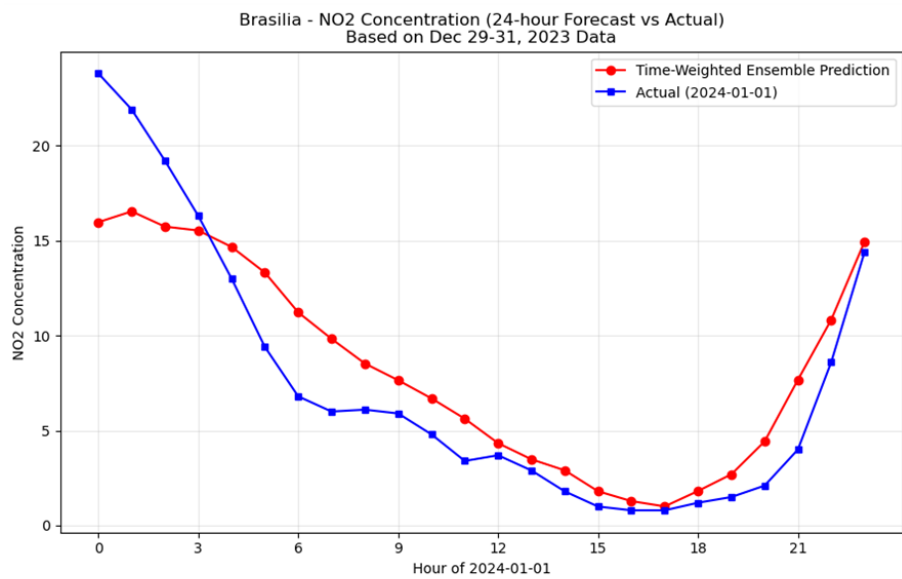


Figure 4. Line chart of CO concentration.

In the predicted results, most of the NO₂ concentrations in Brasilia are around 20 $\mu\text{g}/\text{m}^3$ and the maximum concentration is more than 80 $\mu\text{g}/\text{m}^3$ in March. And the NO₂ concentration of 2024.1.1 is decreasing from 15 $\mu\text{g}/\text{m}^3$ at 0 o'clock and get the lowest point at around 16 o'clock, then increasing to 14 $\mu\text{g}/\text{m}^3$ at the end of the day. Although some values have differences, the two lines decreasing and increasing simultaneously, indicating that the model has good predictive performance. The MAE of this prediction is only 2.25 and the R-squared is up to 0.82.

Afterwards, we draw the comparison graph of 5 models and the actual data. Most of the models can capture the oscillation, the lowest NO₂ concentration of forecast result is almost same as the actual data and the highest one is much lower than the actual data. However, it is easily found that ARIMA has a bad performance in fitting periodicity data, which just captures two oscillation cycles.

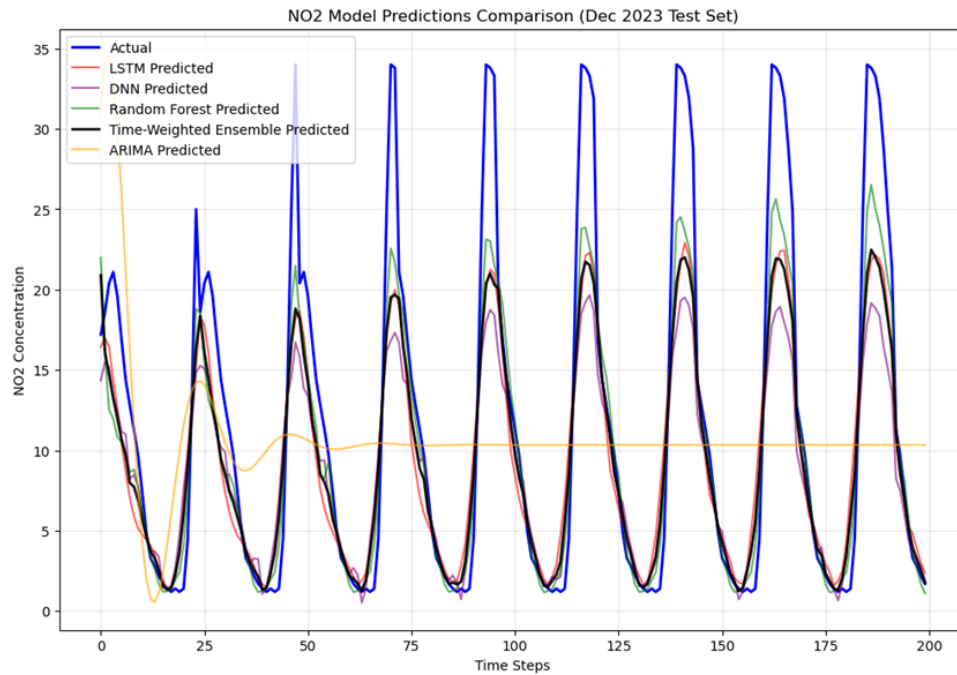


Figure 5. NO₂ model predictions comparison.

5. Model Fit Analysis

In order to evaluate the accuracy and precision of the model prediction results, we use the mean absolute error(MAE) and mean absolute percentage error(MAPE) for measurement.

Mean Absolute Error (MAE) is the average absolute difference between predicted and true values, calculated using the following formula:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Where, y_i is the actual value, $\hat{y}_i$ is the predicted value, N is the number of sample.[17]

The R-squared (R^2), also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. Formula for R-squared (R^2):

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

Where, y_i is the actual value of the dependent variable for the i -th observation, $\hat{y}_i$ is the predicted value of the dependent variable for the i -th observation, $\bar{y}$ is the mean of the actual values of the dependent variable, n is the number of observations.[18]

The following two images show the difference between the predicted value and the true value and the distribution of the prediction error.

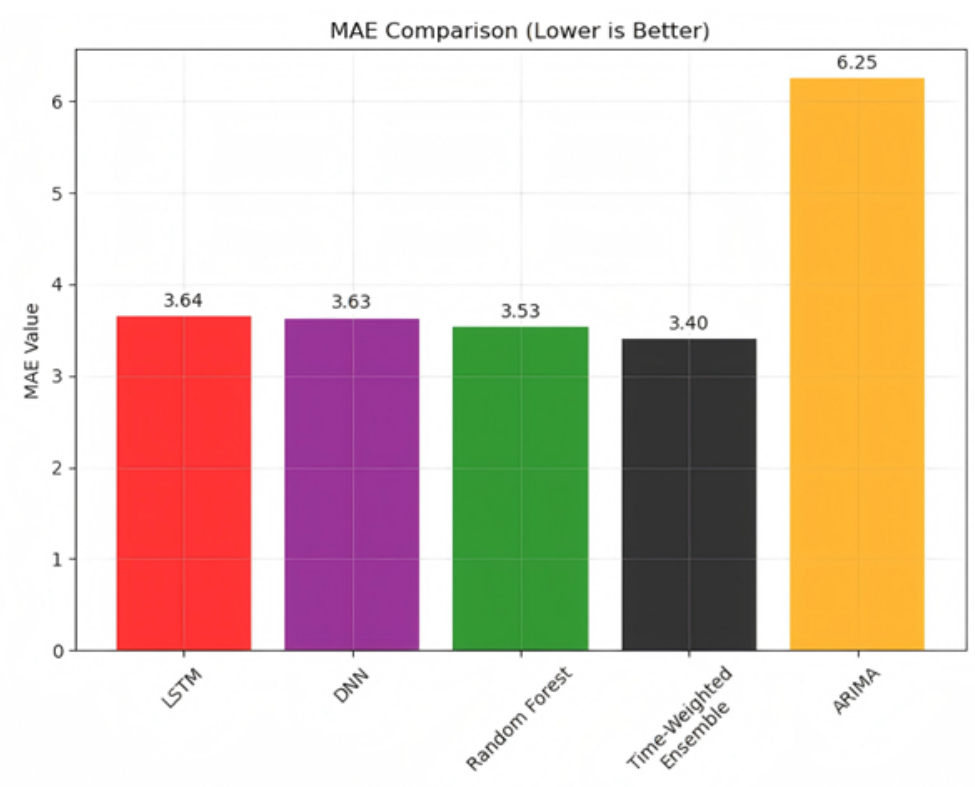


Figure 6. MAE comparison of five models.

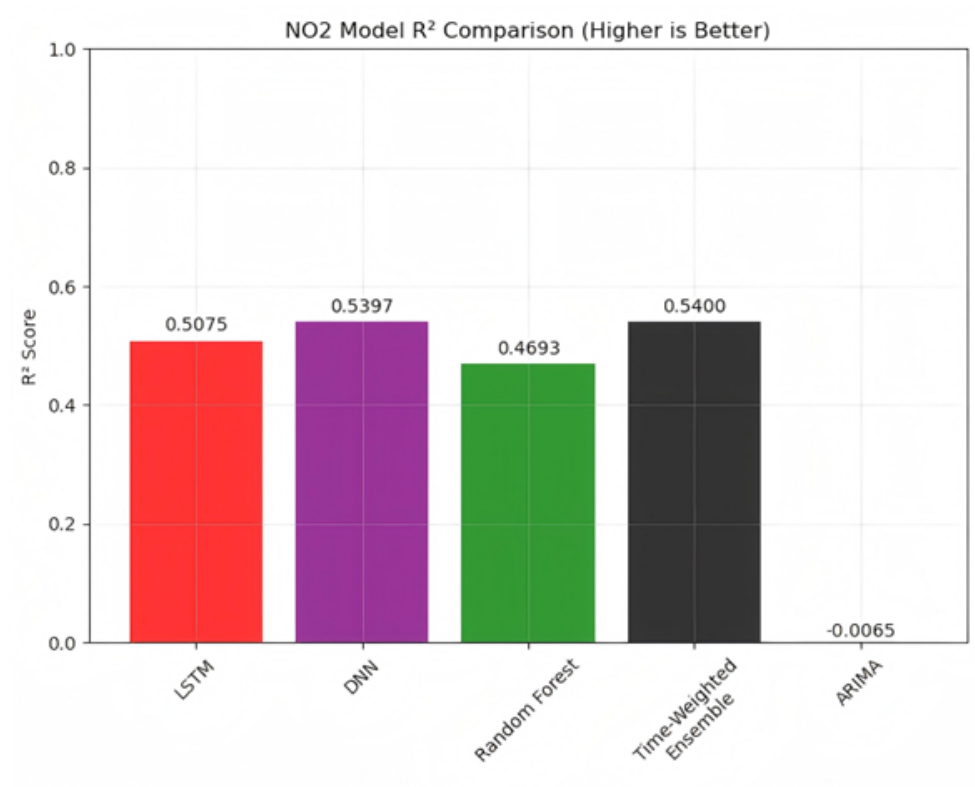


Figure 7. R² comparison of five models.

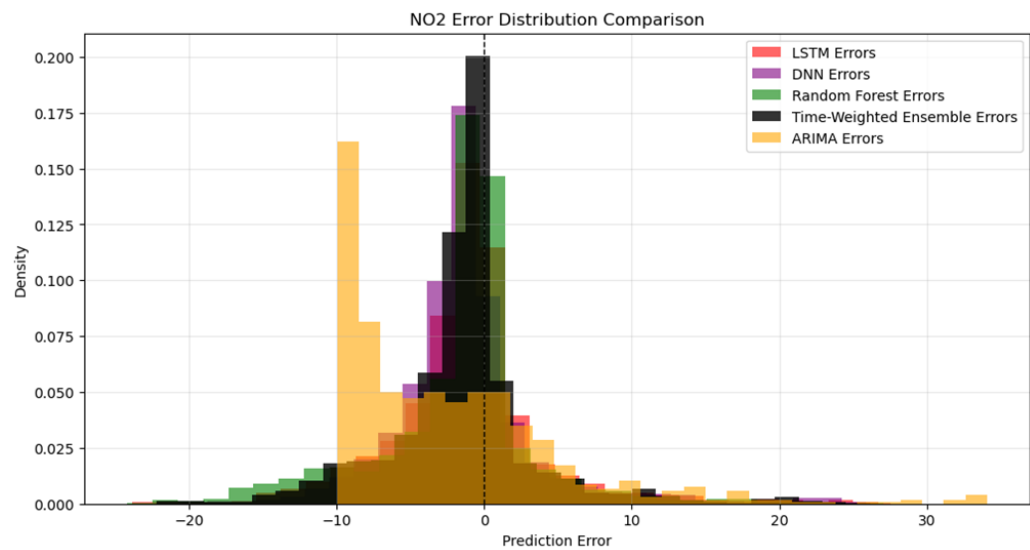


Figure 8. Prediction error distribution.

According to the figure, the MAE of Time-Weighted ensemble model is smaller than other 4 model and the ensemble model also has the best performance in R^2 . Most of the error distribution present a normal distribution, but the ARIMA error distribution get highest density around -10, this is because it did not capture the periodic pattern and get a stable prediction result.

Although, air quality prediction is inherently challenging due to the highly dynamic, nonlinear, and multiscale nature of the atmospheric system, The MAE only 3.40 and the R^2 is 0.54 showing that Time-Weighted ensemble model predicting model still has a good performance in prediction and the error is relatively small.

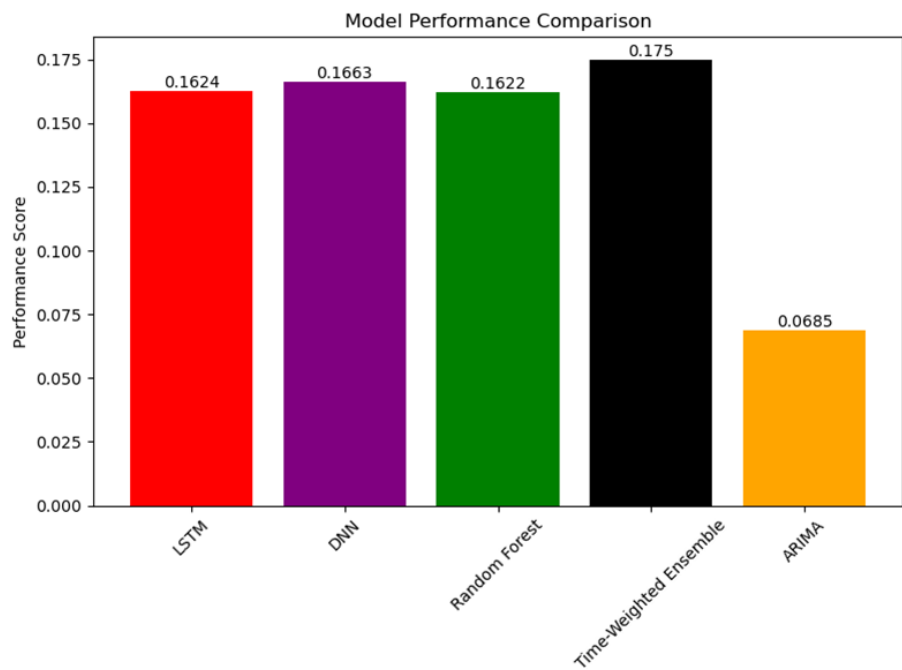


Figure 9. model performance comparison.

We calculated the model performance score by using the performance score function, Time-Weighted ensemble function get the highest score, which improves by 5.23% compared to a single best model.

In addition, we also recorded the model training process and print the training loss function and validation loss function of two neural network model, which are important components of ensemble model.

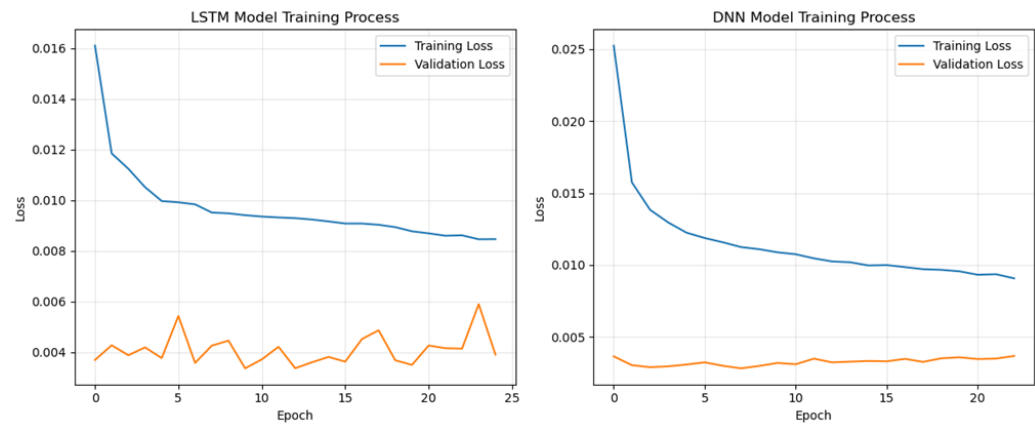


Figure 10. training loss and validation loss of LSTM and DNN.

The training loss of LSTM continuously decreases from around 0.016 to approximately 0.009, indicating that the model is effectively learning and fitting the training data, showing a clear convergence trend. The validation loss remains relatively low and stable, within the range of about 0.004 to 0.005, suggesting that the model maintains a small error on the validation set and demonstrates good generalization ability. And the training loss of DNN also decreases dramatically, from 0.025 to 0.009,the validation loss is always less than 0.005.

6. Establishment of Adaptive Air Quality Model

Adaptive air quality model formula:
For every pollutants $p \in P$, where $P = \{PM_{2.5}, PM_{10}, O_3, NO_2, CO, SO_2\}$, considering that the impact of pollutants on human health is non-linear, we assign an influence coefficient and a safety threshold to each pollutant. After exceeding the safety threshold, I_p will exponentially increase. The health impact values of pollutants I_p is defined as:

$$I_p = \begin{cases} B_p \left[\left(\frac{C_p}{T_p} \right)^{E_p} - 1 \right] & \text{if } C_p > T_p \\ 0 & \text{otherwise} \end{cases}$$

Where, C_p is pollutant concentration, T_p is safety threshold, B_p is basic influence coefficient, E_p is nonlinear index[19]

Table 1. Numerical value of T_p , B_p and E_p .

p	$T_p(\mu \text{ g/m}^3)$ [20]	B_p [21,22]	E_p
PM2.5	15	0.35	1.8
PM10	45	0.25	1.6
O3	50	0.30	2.0
NO2	20	0.20	1.7
CO	230	0.15	1.5
SO2	40	0.10	1.9

Considering the synergistic or antagonistic effects between pollutants, we have designed correlation of pollutants adjustment weight which allocates the weights scientifically:

$$W_p' = W_p \cdot (1 + r_p)$$

$$r_p = \frac{1}{N-1} \sum_{q \neq p} C_{pq}$$

where, W_p is original weight, r_p is average correlation coefficient, N is the number of air pollution, C_{pq} is the correlation coefficient between pollutant p and q in the comprehensive correlation coefficient matrix. W_p' is adjust weight.

Adjusted AAQI value is the weighted sum of all pollutant health impact values:

$$AAQI = \sum_{p \in P} I_p \cdot W'_p$$

Due to the fact that most people are commuting between 7-9 in the morning and 5-7 in the afternoon, and are exposed to more outdoor air pollutants, the impact of air quality on human health is greatest during this period. 10a.m.-16p.m. and 20p.m.-22p.m. are relatively active but generally indoors, while the rest of the time is late at night, mainly during sleep time. They breathe less air and are indoors, so the impact of air quality on human health is minimal during this period. Therefore the time weight factor is defined as:

$$W_t = \begin{cases} 1.4 & \text{if } h \in [7, 9] \text{ and } [17, 19] \\ 1.2 & \text{if } h \in [10, 16] \\ 1.1 & \text{if } h \in [20, 22] \\ 1.0 & \text{else} \end{cases}$$

Where, t represents hour (24 hours per day).

When multiple pollutants coexist, the hazards will be superimposed.[23] We objectively measure the interactions between pollutants using correlation coefficients. The synergistic coefficient is defined as:

$$S = \begin{cases} 1 + 0.5 \cdot \left| \frac{1}{N} \sum_{i=1}^k \sum_{j=i+1}^k r_{p_i p_j} \right| & k \geq 2 \\ 1 & k < 2 \end{cases}$$

Where, k is the quantity of pollutants exceeding the standard ($C_p > T_p$), $r_{p_i p_j}$: correlation coefficient between p_i and p_j , $N = \binom{k}{2} = \frac{k(k-1)}{2}$ is the number of combinations of excessive pollutants.

The difference between the correlation coefficient here and the previous one is that S can be flexibly adjusted according to the different correlation coefficients of pollutants in each city, while the previous only focused on the degree of impact of individual pollutants on health.

The final air quality index is:

$$AAQI_{\text{final}} = AAQI \cdot W_t \cdot S$$

Classification of Air Quality Levels:

Table 2. AAQI range and corresponding level and health proposal .

AAQI range	Level	Health Proposal
[0, 15]	Excellent	Recommmed engaging in outdoor accivities
(15, 30]	Good	Suitable for outdoor activities
(30, 50]	Lightly Polluted	Vulnerable individuals reduce prolonged outdoor activities
(50, 75]	Moderately Polluted	Reduce prolonged outdoor activities
(75, 100]	Heavily Polluted	Avoid outdoor activities
> 100	Hazardous	Stay indoors and use air purifiers

7. Research Result

By calculating the above AAQI model, the AAQI indicators of Brasilia for the whole year of 2023, are shown in the following figure:

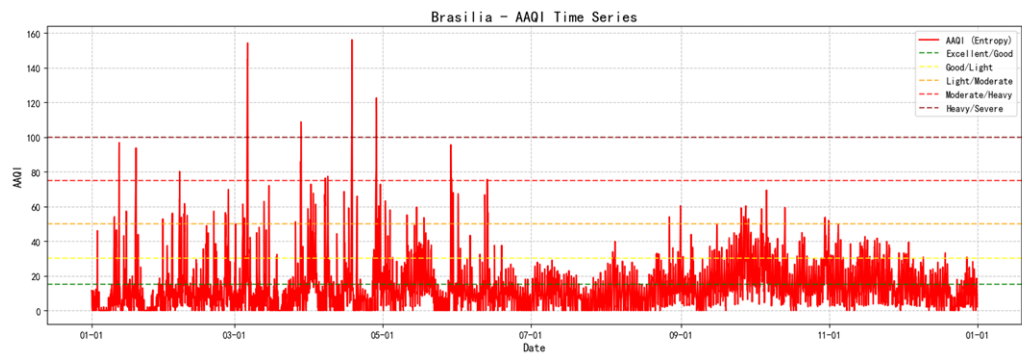


Figure 11. Brasilia 2023 AAQI map.

The average AAQI in Brasilia throughout 2023 is 14.11. Furthermore, The biggest numerical value of AAQI in Brasilia is 160 approximately, few of them are more than 30 and most of them are under 15, which shows that Brasilia has excellent air quality and suitable for people to living. Afterwards, we calculated the average Brasilia 2023 Annual Average Hourly AAQI Index.

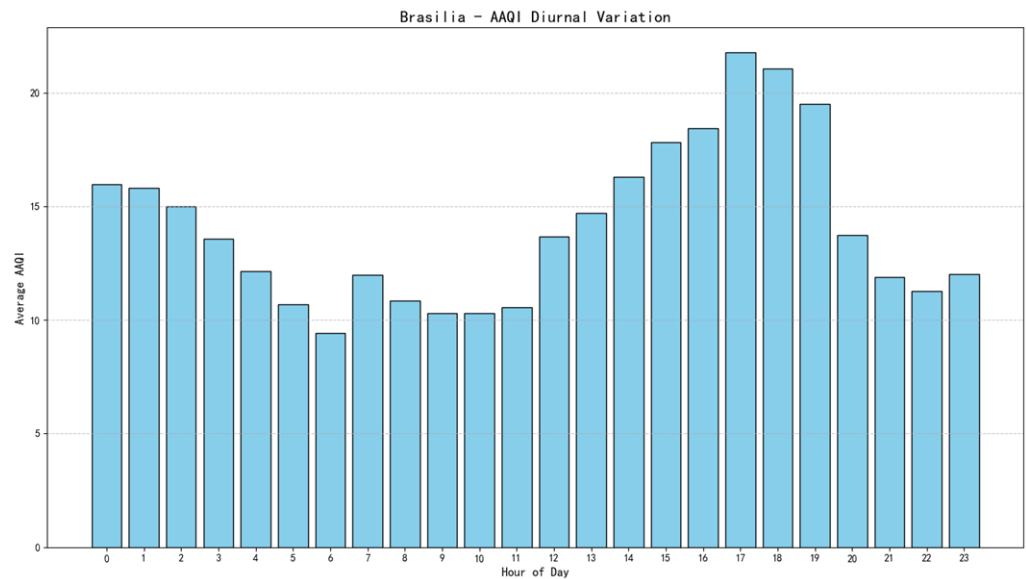


Figure 12. Brasilia 2023 Annual Average Hourly AAQI Index.

This bar chart shows that Brasilia has good air quality around 5 a.m. to 11a.m., and then AAQI gradually increases, reaching its peak at 17 p.m. After that, AAQI rapidly decreases. It presents the air quality tendency daily. This study also calculated the proportion of air quality levels in six cities, namely Brasilia, Cairo, Dubai, London, New York, and Sydney, for the whole year of 2023. The results showed that among these six cities, Sydney had the best air quality, with 72.8% of the days having excellent air quality throughout the year, followed by London, Brasilia, New York, and Cairo. Dubai had the worst air quality, with only 0.1% of the days having excellent air quality throughout the year. It also reminded the governors of Dubai and Cairo should pay more attention to the air pollution control.

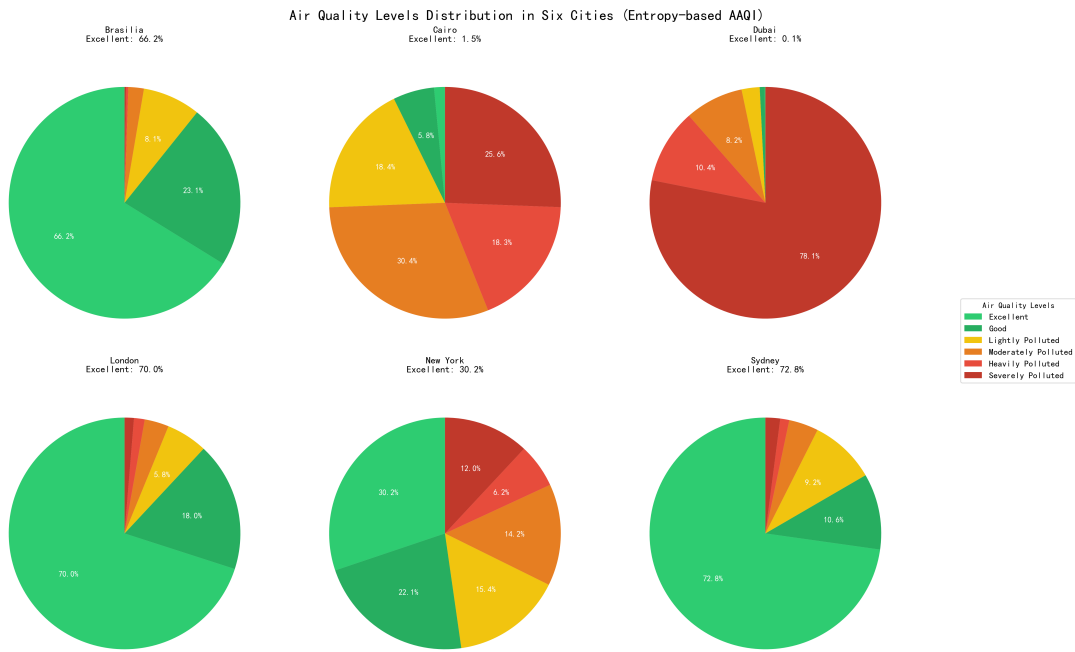


Figure 13. Proportion of Air Quality Levels in Six Cities in 2023.

Among them, the deep green air quality level is excellent, the light green air quality level is good, the yellow air quality level is mild pollution, the orange air quality level is moderate pollution, the red air quality level is severe pollution, and the deep red air quality level is dangerous.

8. Conclusion

This study introduced Time-Weighted ensemble model into the air quality prediction model and achieved good results. The prediction results were consistent with reality, and the R-square of prediction is 0.54, which has 5.23% improvement compared to a single best model. Daily forecast result has a better performance, the mean absolute error is 2.25 and the R-squared is up to 0.82. These results offer a practical tool for issuing timely warnings and helping people reduce exposure to air pollution without many data, and the model can be further improved by adding more data sources or extending to multi-site forecasting. It also created the Adaptive Air Quality Index (AAQI), an advanced model that evaluates air quality from four critical dimensions: concentration, correlation, time, and cooperation. By integrating these factors, the AAQI provides a more comprehensive and dynamic assessment than traditional air quality indices. The model not only considers pollutant levels but also the interactions and temporal variations that influence air quality over time. Beyond being an evaluation system, the AAQI serves as an intelligent link between environmental science and public health. Its innovative framework will help establish future air quality management standards, facilitating a shift from reactive monitoring to proactive prevention in environmental health strategies, ensuring more effective and anticipatory protection of public health.

References

1. warns severe air pollutions, W. World health organization. *Air Quality Guidelines for Europe* **2022**.
2. (outdoor) air pollution, A. World health organization. *World Health Organization* **2024**.
3. Li, X.; Peng, L.; Hu, Y. A hybrid model for short-term air quality index forecasting based on mode decomposition and deep learning. *Atmospheric Environment* **2021**, *246*, 118096.
4. Priti, K.; Kumar, P. A critical evaluation of air quality index models (1960–2021). *Environmental Monitoring and Assessment* **2022**, *194*, –.
5. Kaggle. Youssef Elebiary. *Global Air Quality (2023) - 6 Cities[Data set]* **2024**.
6. Kaggle. Youssef Elebiary. *Global Air Quality (2024) - 6 Cities[Data set]* **2025**.

7. Taunk, K.; De, S.; Verma, S.; Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. 2019 Int Conf Intell Comput Control Syst ICCS 2019, no. May 2019, 2019.
8. Seinfeld, J.H.; Pandis, S.N. *Atmospheric chemistry and physics: from air pollution to climate change*; John Wiley & Sons, 2016.
9. Pedregosa Fabianpedregosa, F.; Michel, V.; Grisel Oliviergrisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-learn: machine learning in Python.
10. Shannon, C.E. A mathematical theory of communication. *The Bell system technical journal* **1948**, *27*, 379–423.
11. Spearman, C. The proof and measurement of association between two things. **1961**.
12. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93.
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
14. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.
15. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
16. Box, G. Box and Jenkins: time series analysis, forecasting and control. In *A Very British Affair: Six Britons and the Development of Time Series Analysis During the 20th Century*; Springer, 2013; pp. 161–215.
17. Draper, N.R.; Smith, H. *Applied regression analysis*; Vol. 326, John Wiley & Sons, 1998.
18. Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W.; et al. *Applied linear statistical models* **1996**.
19. Miller, N.H.; Molitor, D.; Zou, E. The nonlinear effects of air pollution on health: Evidence from wildfire smoke. Technical report, National Bureau of Economic Research, 2024.
20. World Health Organization (WHO). Air quality guidelines: Global update 2021. 2021. Available online: <https://www.who.int/publications/i/item/9789240034228>. <https://doi.org/10.1287/trsc.1040.0107>
21. Huang, S.; Li, H.; Wang, M.; Qian, Y.; Steenland, K.; Caudle, W.; Liu, Y.; Sarnat, J.; Papatheodorou, S.; Shi, L. Health impacts of long-term exposure to ambient air pollution: A global systematic review and meta-analysis. *Science of the Total Environment* **2021**, *776*, 145968.
22. Pope III, C.A.; Dockery, D.W. Health effects of fine particulate air pollution: lines that connect. *Journal of the air & waste management association* **2006**, *56*, 709–742.
23. Xia, L.; Zhou, S.; Han, L.; Sun, W.; Sun, H. Joint association of air pollutants on cardiometabolic multimorbidity. *Scientific Reports* **2024**, *14*, 26987.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.