Prepublished copy

Nature Environment & Pollution Technology

This is a peer-reviewed prepublished version of the paper

Type of the Paper (Original Research)

A Comparative Study of Statistical and Machine Learning Modelling Techniques in Air Pollution Data

Sumithra Palraj[1†], Loganathan Appaia, Deneshkumar Venugopal[3] and Gunasekaran Munian[4]

[1] Department of Statistics, Pondicherry University, Pondicherry, 605 001; 350433@msuniv.ac.in
[2] Department of Statistics, Central University of Tamil Nadu, Thiruvarur, India-610005; loganathanappaia@gmail.com
[3] Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, India-627012; vdenesh77@gmail.com
[4] Department of Applied Mathematics, The Gandhigram Rural Institute, Institute (Deemed to be University) Dindigul, India- 624 302; gunasekaranm1988@gmail.com
† Corresponding author: Sumithra Palraj; 350433@msuniv.ac.in
0000-0002-3800-6024

Key Words	Autoregressive moving average, Facebook prophet, Long short-term memory, Air Quality Index
DOI	https://doi.org/10.46488/NEPT.2025.v24i04.B4298 (DOI will be active only after the final publication of the paper)
Citation for the Paper	Palraj, S., Appaia, L., Venugopal, D. and Munian, G. 2025. A Comparative Study of Statistical and Machine Learning Modelling Techniques in Air Pollution Data. *Nature Environment and Pollution Technology*, 24(4), p. B4298. https://doi.org/10.46488/NEPT.2025.v24i04.B4298

Abstract: Different approaches are being adopted, in practice, for determining models for given time series. The approaches can be categorized broadly into three viz., statistical, machine learning and deep learning. Since they differ with respect to their theoretical base, their outcomes also differ. Decision making based on the values predicted from the time series models seek accuracy of the forecast values. This paper studies the effectiveness of the three approaches by comparing the performance of the autoregressive moving average method developed applying statistical principles, Facebook Prophet method developed from Machine Learning approach and long short-term memory method developed from deep learning. The study is carried for real data of time series of air quality indices.

1. INTRODUCTION

One of the main applications of time series models is forecasting values with desired level of accuracy based on the observations available for the past time periods. This is crucial for decision making in various fields, as it allows for the prediction of future trends and patterns and the identification of potential risks and opportunities. Buyuksahin and Ertekin (2019) have mentioned that time series models can also be used to extract meaningful information from large amount of data and to identify underlying relationships and patterns in the data. In recent years, there has been a significant amount of research on developing methods for time series modelling applying machine learning and deep learning approaches. These methods have the ability to handle large, complex datasets. Parmezan *et. al.*, (2019) have pointed out that the models developed from these approaches can identify patterns and trends, which may be difficult to discern using conventional statistical methods.

It is well known that the autoregressive moving average / autoregressive integrated moving average (ARIMA/ARMA) methods have been developed based on statistical principles (Box *et. al.*, (1970)) and the outcomes of the models possess desirable theoretical properties. Machine learning techniques, such as support vector machine (SVM) and random forests have also been shown effective in forecasting time series data. Deep learning methods, such as long short-term memory (LSTM) method, have demonstrated the ability to handle large amount of data and capture complex relationships in the data (Toharudin *et. al*., (2020)).

Current research in time series modelling has also focused on improving the accuracy and efficiency of forecasting methods. Calvo and Santafe Rodrigo (2016) have mentioned that different time series modelling methods can be compared with respect to various features such as seasonality, trend, level, noise, *etc*., of the time series data, ability to handle different types of data, computational complexity of the methods, as well as the interpretability of the results and accuracy of the forecast values.

In recent years, air pollution has become a major concern for cities and communities around the world. Increasing levels of toxic substances in the air have caused a wide range of health problems from respiratory issues to heart diseases. The most common sources for air pollution include industrial emissions, transportation and energy production. As a result, the common public are breathing polluted air every day, leading to a rise in respiratory diseases and other illness. Though industrial growth leads to economic development of nation, it parallelly returns poor quality air to the humankind. Hence, there is a considerable amount of consideration worldwide for discussing Climate Change. Analysis of information about contents of air, which all human beings inhale, is pertinent in view of health of common public. In recent years, several studies have examined the use of ARIMA models for forecasting air pollutant concentrations.

Samia *et. al.*, (2012) conducted a study to predict the levels of PM_{10} over 24 hours period in Tunisia. They proposed a hybrid method, which combines the ARIMAX and artificial neural network (ANN) models with a multilayer perceptron architecture. This model was able to identify both linear and non-linear time series patterns effectively, making it a valuable tool for forecasting and early warning systems**.** Haviluddin *et.al.*, (2015) compared the performance of the ARIMA method, Back-Propagation Neural Network (BPNN) and genetic algorithms for analysing and predicting short-term time series network traffic activity datasets. They experienced that BPNN is efficient in learning a time series data. Naveen and Anu (2017) compared the effectiveness of ARIMA and Seasonal ARIMA models for forecasting ambient air quality data in Thiruvananthapuram District, Kerala, India, and found that the ARIMA model performed better. Patra (2017) used a combination of ARIMA method and artificial intelligence techniques like ANN and SVM to predict the concentration of atmospheric pollutants in India, and obtained satisfactory results with the ARIMA model.
Abhilash *et. al*., (2018) found that the ARIMA model is effective for short-term forecasting of *nitrogen dioxide* (NO_2), PM_{10} and s*ulphur dioxide* (SO_2) for the data pertaining to Bengaluru city.

Parmezan *et. al.*, (2019) compared the effectiveness of the four algorithms *viz*., ANN, SVM, *k*NN-TSPI and SARIMA in predicting trends in multi-step-ahead projections with approximate or updated iterations. Their results showed that the *k*NN-TSPI method has notable stability and robust in forecasting as accurately as SARIMA and SVM, but with the added advantage of being easier to understand, implement, and modify with only two input parameters that can be easily determined based on seasonal patterns in the data. These findings suggest that machine learning methods have reached a similar level of maturity to statistical models in temporal data modelling.

In a study by Samal *et. al*., (2019) the SARIMA and Prophet models were evaluated for predicting pollution in Bhubaneswar City using historical data. These methods were found, based on various performance measures, to achieve high levels of accuracy. Particularly, the Prophet model with *log* transformation was found to be relatively more effective for forecasting future pollution levels in Bhubaneswar City.

In the study by Toharudin *et. al*., (2020), the LSTM and the Prophet models were used to analyze historical data of daily maximum and minimum air temperatures in Bandung. Their findings showed that the Prophet model performs relatively better for forecasting maximum air temperature, while the LSTM model performs relatively better for forecasting minimum air temperature. Sanchez-Pozo *et. al*., (2021) compared the accuracy in the forecast values, of Ozone (O_3) concentration for assessing air quality of London city, obtained from the four forecasting models Facebook Prophet (Prophet), LSTM, Support Vector Regression (SVR) and ARIMA. They observed that LSTM model achieved the highest accuracy among the models considered. Some works could be found in the literature on studying the quality of air and chemical pollutants causing air pollution in Chennai, a metropolitan city in India. Nadeem *et. al.*, (2020), applying the ARIMA method, developed nine univariate linear stochastic models for forecasting the concentration levels of the chemical pollutants *viz*., *respirable suspended particulate matter*, SO_2 and NO_2 in three areas of Chennai city. Ammasi Krishnan *et. al*., (2020) found that the level of $PM_{2.5}$ is three times higher than the National Ambient Air Quality Standard (NAAQS) of India while other pollutants are within acceptable limits, during *Diwali* days.

Angelena *et. al*., (2021) examined the concentration of PM_{10} in the four locations Anna Nagar, Adyar, Thiyagaraya Nagar and Kilpauk of Chennai city applying a neural network with multilayer feed-forward backpropagation method. Whereas, Janarthanan *et. al*., (2021) found that a combination of the deep learning model based on LSTM and SVR could accurately predict well the air quality index (AQI) of some locations of Chennai. Mani and Volety (2021), applying the LSTM and ARIMA methods, predicted the levels of air pollutants and found that the LSTM model produced relatively precise results than the ARIMA model.

Varma *et. al*., (2021) investigated and analysed the concentration of *suspended PM* and its associated health consequences in five locations Alandur, Marina Beach Road, Velachery, Koyambedu and Anna Salai of Chennai city. They discovered that out of the five locations, Alandur, Marina Beach Road and Velachery have moderate AQI values, while Koyambedu and Anna Salai have poor AQI values. Recently, Gunasekar *et. al*., (2022) proposed a novel and optimised deep learning algorithm, which combines ARIMA and CNN - LSTM methods for forecasting the AQI of Chennai with higher level of accuracy.

Apart from time series analysis, some attempts are also found in literature on model construction for AQI of Chennai city. Sumithra *et. al*., (2022) have discussed Markov modelling for AQI of Velachery locations of Chennai. In another work, Loganathan *et. al*., (2022*a*) have constructed multiple regression model for the same datasets.

It can be noted, in general, from the above discussions that performance of the approaches or methods applied for time series modelling are data dependent. AQI time series is challenging to handle due to its inherent complexity, non-stationarity and susceptibility to external factors such as weather conditions, seasonal variations and human activities. These factors cause patterns to shift unpredictably, making it difficult to model and forecast accurately. Unlike other types of data, air pollution data often exhibits irregular trends, sudden spikes, and long-term dependencies, which require specialized techniques to capture and analyze effectivel. Addressing these patterns is crucial because accurate forecasting of air pollution levels enables proactive measures to mitigate health risks, inform public policy and improve environmental management.

Hence, it has become essential to identify appropriate approach or method for designing effective forecasting models for every location. The main objective of this work is to identify an effective method for forecasting the AQI of Chennai city on comparing the performance of three methods *viz*., ARIMA, Prophet and LSTM evolved respectively from statistical, machine learning and deep learning approaches. Implements these three distinct methodologies for time series analysis: ARIMA, LSTM and Prophet. Each methodology was selected based on unique capabilities in addressing specific aspects of time series analysis and forecasting challenges.

The ARIMA methodology demonstrates proven effectiveness in modeling linear relationships and handling datasets with distinct trends and seasonality patterns. This approach combines autoregressive and moving average components, utilizing differencing techniques to achieve data stationarity. Through systematic parameter tuning and iterative testing processes, the model specifications were established for accurate forecasting capabilities.

For addressing non-linear and complex patterns, the LSTM architecture leverages advanced capabilities in modeling long-term dependencies within sequential data. The network architecture incorporates multiple layers with carefully configured neurons across input, hidden and output layers. The implementation includes optimized hyperparameters and activation functions, ensuring effective information flow and gradient management throughout the training process while maintaining computational efficiency and preventing model overfitting.

Facebook Prophet methodology addresses multiple seasonal patterns and irregular events within the time series data. This sophisticated approach incorporates automated detection mechanisms for various seasonality patterns while effectively managing holiday effects and outlier scenarios. The model's configuration includes robust outlier handling mechanisms and flexible seasonality modeling, providing significant advantages over traditional forecasting approaches.

During COVID-19 Lockdown, the air pollution levels were significantly lower. The primary reason for this reduction was the drastic decrease in vehicular movement due to strict lockdown measures. With fewer vehicles on the road, industrial activity also slowed down, leading to minimal fluctuations in air pollution data.

Since time series models rely on variations in data to make accurate predictions, the lockdown period, characterized by consistently low pollution levels, does not provide meaningful insights for model comparison. Instead, we focused on the Before and After COVID-19 Lockdown periods, where pollution levels exhibited notable variations due to human and economic activities(Laxmipriya and Narayanan, 2021).

Performance of the three time series modelling methods is compared based on the AQI data of three monitoring stations in Chennai city Velachery, Alandur and Manali. Section 2 presents an overview of the study area and statistical descriptions of the data. The methodology followed in the three methods and the performance measures considered for comparison are outlined in Section 3. Time series models are constructed based on the AQI values collected from each of the three monitoring stations applying the three model building methods in Section 4. Performance of the models are also compared for each monitoring station for before and after COVID 19 lockdown periods. Section 5 presents the concluding remarks.

2. STUDY AREA AND DATA DESCRIPTIONS

2.1 Study Area

Chennai is one of the four major metropolitan cities in India and is the capital of Tamil Nadu state. The population size of this city is increasing mainly due to the growth of the manufacturing and Information Technology industries, which is leading to an influx of people from other cities and neighbouring states. As reported by Tamil Nadu Pollution Control Board, 2021 (Sekar *et.al.* 2020) the population density of the city is 26,553 people *per square kilometre*. The Central Pollution Control Board (CPCB) of India operates three continuous ambient air quality monitoring stations in the city at Velachery, Alandur and Manali. Velachery is a commercial and residential area located in the southern part of Chennai, Alandur is a residential area in the southwest of the city and Manali is an industrial area located in the northern part of Chennai. These stations record daily data on levels of pollutants for the respective region such as *PM*2.5, *NO*, NO_2, *NOx*, NH_3, SO_3, O_3, *Benzene* and *Toluene* as well as meteorological factors such as wind speed, direction, relative humidity and temperature. In each station, as per the guidelines of CPCB, the sub-indices are computed every day for each chemical pollutant and the highest sub-index is determined as the value of AQI of the region.

2.2 Data Descriptions

The AQI data was collected from Velachery monitoring station for a period of 813 days from January 1, 2018 to April 24, 2020; from Alandur monitoring station for a period of 1178 days from January 1, 2017 to March 24, 2020; and from Manali monitoring station for a period of 479 days from December 1, 2018 to April 24, 2020. All these time series contain missing values and outliers. Occurrence of missing observations in a data is not uncommon. It is due to the missing records caused by phenomena, faulty equipment, loss of records or a mistake that cannot be rectified. When one or more observations are missing, it may be necessary to replace them by suitable estimates. Such replacements can enable to provide better understanding of the nature of the data. Some of the standard ways of handling missing data include deleting and imputation (The missing values are either left out or replaced with a single substitute). In time series, the temporal dependency is essential and deleting values may interrupt this continuity and replacement methods may change the original time series, causing reduced statistical power and biased estimations. Missing values in a time series can make it difficult to determine correlations with past lag (Huang *et. al.*, (2018)) and presence of outliers in the data may lead to biased forecasting result. This can be overcome by replacement of appropriate values in the positions of missing observations and by suitable estimates in the positions of outliers.

Kihoro *et. al.*, (2013)) have suggested that model-based estimation methods can be more statistical, which can be followed to determine such values. Here, the LSTM model-based method is applied to determine such replacement values and estimates. More details can be found in Loganathan *et. al.*, (2023 b).

Plot of the AQI data for the three locations - Velachery, Alandur and Manali - are displayed in Figure 1. The figure enables to identify patterns and trends in the AQI data over time for each location. The figure reveals that the plot of AQI of Velachery has spikes in June, October and November of 2018, followed by a decreasing trend in 2019. The plot of AQI of Alandur exhibits an upward trend in AQI at the end of each year, with relatively little fluctuations during other periods. The plot of AQI of Manali shows a moderate downward trend in AQI over the years.

Table 1 presents statistical information about AQI of the three different locations. The levels of AQI in Velachery range from 17.50 to 194.09 with a mean value of 63.77 and standard deviation of 29.69. Twenty-five percent of the values of AQI in Velachery lie below 41.00; half of the information lies below/above the median value of 57.05 and remaining twenty-five percent of the values lie above 79.72. From the co-efficient of skewness and co-efficient of kurtosis, respectively 1.23 and 1.70, it can be noted that the distribution for AQI could be positively skewed and platykurtic.

The levels of AQI in Alandur range from 19.50 to 378.19 with a mean value of 83.70 and standard deviation of 39.01. Twenty-five percent of the values of AQI in Alandur lie below 60.11; half of the information lies below/above the median value of 76.02 and remaining twenty-five percent of the values lie above 95.07. The co-efficient of skewness and coefficient of kurtosis, respectively 1.55 and 4.15, point out that the distribution for AQI could be positively skewed and leptokurtic.

The levels of AQI in Manali range from 16.54 to 351.61 with a mean value of 117.10 and standard deviation of 66.04. Twenty-five percent of the values of AQI in Manali lie below 70.62; half of the information lies below/above the median value of 96.08 and remaining twenty-five percent of the values lie above 145.00. The co-efficient of skewness and coefficient of kurtosis are 1.53 and 2.01respectively, which suggest that the distribution for AQI in Manali could be positively skewed and leptokurtic.

The coefficient of variation for AQI of Velachery, Alandur and Manali are 47%, 47% and 56% respectively. These indicate that consistency in the time series of AQI of Velachery and Alandur are equal and are relatively more consistent than the time series of AQI of Manali.

According to the categorization of National Air Quality Index consideration range (https://app.cpcbccr.com/ccr/#/login), the average air quality in Velachery and Alandur is *Satisfactory* and in Manali is *Moderately* polluted.

Table 1: Descriptive Statistical Measures

Statistical Measure	Velachery	Alandur	Manali
Minimum	17.50	19.50	16.54
Maximum	194.09	378.19	351.61
Mean	63.77	83.70	117.10
Median	57.05	76.02	96.08
Standard Deviation	29.69	39.01	66.04
Quartile 1	41.00	60.11	70.62
Quartile 3	79.72	95.07	145.00
Co-efficient of Variation	0.47	0.47	0.56
Co-efficient of Skewness	1.23	1.55	1.53
Co-efficient of Kurtosis	1.70	4.15	2.01

Fig. 1: AQI of Velachery, Alandur and Manali Stations

3. METHODOLOGY

3.1 Autoregressive Moving Average Method

The ARMA model is estimated for a given time series applying the Box-Jenkins approach to predict the values for future time periods (Box *et. al*., (1970)). The general form of ARMA model is given by

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + e + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots + \phi_q \varepsilon_{t-q}$$

where

y_t is the observation of time t

α is the intercept

β_i is the coefficient of autoregressive (AR) component with lag *t-i*, *i*=1, 2, 3, …, *p*

ϕ_j is the coefficient of moving average (MA) component with lag *t-j*, *j*= 1, 2, 3, …, *q*.

Determination of the coefficients α, β_i (*i*=1, 2, 3, …, *p*) and ϕ_j (*j*= 1, 2, 3, …, *q*) for the given time series is popularly knowns as fitting of the ARMA(*p*, *q*) model. Various unit root tests are applied for evaluating the stationarity of a given time series statistically, before determining the estimates. Dickey-Fuller test is a commonly used test of this type, which tests

H_{01}: Time series is not stationary

against

H_{11}: Time series is stationary.

If the value of the test statistic is significant, the series can be considered as stationary. Ljung – Box test is applied to determine if there is any significant autocorrelation in the given time series, which tests the hypotheses,

H_{02}: Residuals are independently distributed

against

H_{12}: Residuals are not independently distributed.

Optimal values of p and q for the ARMA model can be determined from the values of AIC and BIC. It is generally recommended to choose the combination of p and q corresponding to the relatively lower values of AIC or BIC. Plot of the values of autocorrelation function (ACF) and partial autocorrelation function (PACF) are also considered for diagnosing the values of p and q.

3.2 Facebook Prophet Method

The Facebook Prophet model decomposes a time series into three components *viz.*, trend, seasonality and holidays. The trend component captures the overall direction and level of the time series, while the seasonality component captures periodic fluctuations in the data. The holiday component is used to model the effect of special events or holidays in the time series data. Together, these three components can be used to model and forecast future values of the time series. The general form of the Facebook Prophet model is given by (Taylor and Letham, 2018)

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

where

$y(t)$ represents the time dependent response variable,

$g(t)$ is the trend function,

$s(t)$ is the seasonality component,

$h(t)$ represents the holiday effect, and

ϵ_t is the measure of overall uncertainty in the model.

By choosing changepoints from the data, a piecewise linear trend automatically finds changes in trends and its form is

$$g(t) = \left[\left(k + \sum_{i \geq t} \delta_i\right) t + m + \sum_{i \geq t} (-change\ point_i \delta_i)\right] * y_{scaled}$$

where k is the growth rate, δ is the rate adjustment, m is the offset parameter and y_{scaled} is the time series data that has been scaled (normalized) to have values between 0 and 1. This is done to make the data easier to work with and to ensure that the trends and seasonality are not affected by the scale of the data.

The seasonality component is modelled using the Fourier series as

$$s(t) = \sum_{j=1}^{n} \left(a_n cos\left(\frac{2\pi jt}{P}\right) + b_n sin\left(\frac{2\pi jt}{P}\right)\right)$$

where P is the regular period determining the periodic fluctuations in the data. A high value of P indicates strong seasonality, while a low value indicates weak seasonality.

The holiday component $h(t)$ refers to specific days of the year that deviate from the normal schedule or routine, such as holidays or special events. These days are predictable and they occur at the same time every year.

In the general form, ϵ_t represents the uncertainty or error in the model prediction for time *t*. It is used to model the variability or noise in the data, which the model is trying to predict. It is important to note that the value of ϵ_t will vary over time, depending on the data being used and the model performance. Taylor and Letham (2018) have mentioned that a low value of ϵ_t indicates more accurate prediction and high value indicates greater uncertainty.

3.3 Long Short-Term Memory Method

LSTM method was proposed by Henreireu (1976), which is one of the methods developed applying the deep learning approach for fitting linear and non- linear stationary time series models. The LSTM model can identify long-range patterns in the time series.

As described by Van Houdt *et. al.*, (2020), the LSTM method functions with three gates *viz.*, input gate, forget gate and output gate. This method functions in the three gates with a set of equations as explained below: The LSTM unit gets the current input sequence, given by x^t, and the output from the preceding time step, denoted by h^{t-1}. The weighted inputs are added together and then transferred through *tanh* activation to produce $\widetilde{C_t}$ as

$$\widetilde{C_t} = tanh(W_c x^t + U_c h^{t-1} + b_c)$$

The input gate directs the flow of data into the memory cell, through the *sigmoid* activation function

$$i_t = \sigma(W_i x^t + U_i h^{t-1} + b_i)$$

The forget gate reads x^t and h^{t-1} and activates weighted inputs with the *sigmoid* function. The forget gate f_t is multiplied by the cell state at the preceding time step C^{t-1}, allowing the memory contents, which are no longer considered as forgotten.

$$f_t = \sigma\left(W_f x^t + U_f h^{t-1} + b_f\right)$$

The output gate uses *sigmoid* activation function to regulate the information flows out of the LSTM unit by taking the weighted sum of x^t and h^{t-1}.

$$O_t = \sigma(W_O x^t + U_O h^{t-1} + b_O)$$

The current cell state C_t is calculated by ignoring irrelevant information from the preceding time step and accepting relevant information from the current input. This is mainly composed of the constant error carousel, which has a recurring edge with unit weight.

$$C_t = \widetilde{C_t} * i_t + C_{t-1} * f_t$$

By transferring the cell state C_t through a *tanh* function and multiplying i_t with the output gate O_t, the LSTM unit output, h_t is calculated from

$$h_t = \tanh(C_t) * O_t$$

where

$x^t \in \mathbb{R}^d$ is the input vector for the LSTM unit,

$f_t \in \mathbb{R}^h$ is the forget gate activation vector,

$i_t \in \mathbb{R}^h$ is the input / update gate activation vector,

$O_t \in \mathbb{R}^h$ is the output gate activation vector,

$h^t \in \mathbb{R}^h$ is the hidden state vector,

$\widetilde{C_t} \in \mathbb{R}^h$ is the cell input activation vector,

$C_t \in \mathbb{R}^h$ is the cell state vector and

$W \in \mathbb{R}^{h*d}$, $U \in \mathbb{R}^{h*h}$ and $b \in \mathbb{R}^h$ are respectively the input weight, recurrent weight and bias vector.

Here, *d* and *h* refer to the number of input features and number of hidden units respectively and σ is the *sigmoid* activation function.

The LSTM network architecture was carefully designed based on our data characteristics and experimental validation, with key architectural decisions focused on balancing complexity and efficiency. A three-layer structure was implemented, beginning with an input layer of 50 neurons to process initial data, followed by a hidden layer of 50 neurons to capture intermediate patterns, and concluding with an output layer containing neurons aligned with the prediction horizon. Hyperparameters were rigorously justified to optimize performance: the memory cells integrated three sigmoid layers and one tanh layer to regulate information flow, while a learning rate of 0.001 ensured stable convergence during training. A batch size of 32 samples was selected to balance computational efficiency and gradient stability, and a dropout rate of 0.2 was applied to minimize overfitting risks. These design choices collectively aimed to enhance the model's predictive accuracy while maintaining computational practicality.

3.4 Model Evaluation Measures

The *train - test* split of the time series is implemented to assess the model's effectiveness with unseen data. By training the model on a portion of the data and evaluating its performance on a separate portion with test data part, a better understanding can be gained on how the model will perform for other datasets. If the model found performing well on the test data, confidence can be placed on the ability of the model to perform well on future data(Briggs *et al.*, 1999).

In the present study, the AQI recorded in the three monitoring stations are split into training (95%) and test (5%) data. In Velachery monitoring station, the AQI from 01-01-2018 to 11-02-2020 are considered as the *training data* and the AQI from 12-02-2020 to 23-03-2020 are considered as the *test data*. For Alandur monitoring station, the AQI from 01-01-2017 to 24-02-2020 are considered as the *training data* and the AQI from 25-01-2020 to 23-03-2020 are considered as the *test data*. For Manali monitoring station, the AQI from 12-01-2018 to 28-02-2020 are considered as the *training data* and the AQI from 29-02-2020 to 23-03-2020 are considered as the *test data.*

The 5% test split, though small in proportion, represents over one month of contiguous daily data, ensuring sufficient temporal coverage for statistically meaningful evaluation. Forecasting beyond one month in daily time series introduces significant challenges due to error accumulation and increased uncertainty, as small deviations compound over time, reducing prediction reliability. The test period maintains a balance between robust validation and minimizing the risk of error propagation, ensuring the model's performance metrics remain interpretable and representative of real-world forecasting scenarios(Briggs *et al.*, 1999).

Toharudin *et al*. (2020) have stated that mean absolute error (MAE) root mean squared error (RMSE) and mean absolute percentage error(MAPE) are commonly used measures to assess a model's predicting capability. These measures are used here to evaluate the model performance and to make model comparison(Zhang *et al.*, 2013). Values of these measures can be calculated for a given time series using the following formulae.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \widehat{Y}_t\right|$$

$$\text{MAPE}=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t-\widehat{Y}_t}{Y_t}\right|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \widehat{Y}_t\right)^2}$$

Where,

Y_t is the actual value at time t.
$\widehat{Y_t}$ is the estimated value at time t.
n is the number of observation.

The MAE measures the average magnitude of the errors in a model's predictions, whereas RMSE considers the average squared difference between the predicted and actual values. MAPE provides a relative measure, which is useful for comparing across datasets with different scales.

4. TIME SERIES MODELS AND EVALUATION

4.1 Time Series Models for AQI of Velachery

4.1.1ARMA Model for Velachery

The stationarity of the time series of AQI recorded in Velachery monitoring station is tested applying the Dickey-Fuller test. The *p*-value of the test statistic is computed as 0.0019, which indicates that the time series of AQI recorded in Velachery monitoring station can be stationary. The order of the ARMA model for this time series can be diagnosed from the ACF and PACF plots displayed in Figure 2.

The ACF plot shows that the first 20 *lags* are significant and the PACF plot indicates that the lags 1, 5 and 13 are significant. The ARMA models are estimated corresponding to each pair (*p*, *q*) with *p* = (1, 2, 3, 13) and *q* = (1, 2, 3). The values of AIC and BIC are calculated for each ARMA model and are displayed in Table 2.

It can be noted from Table 2 that the smallest value of BIC, 159.11, is obtained corresponding to the ARMA (2,1) model, whereas the smallest value of AIC, 134.34, is obtained corresponding to the ARMA (3,2) model. The *p*-values calculated for the construction of the ARMA (2,1) model reveal that the model is not significant to the time series considered. The results corresponding to the construction of the ARMA (3,2) model are presented in Table 3.

Table 2: AIC and BIC Values for ARMA Models of Velachery Station

S. No	*p*	*q*	AIC	BIC
1	1	1	147.468	166.064
2	1	2	148.380	171.625
3	1	3	146.436	174.330
4	2	1	135.86	159.11
5	2	2	136.74	164.64
6	2	3	137.47	170.02
7	3	1	136.59	164.48
8	3	2	134.34	166.88
9	3	3	136.17	173.36
10	13	1	145.31	219.69
11	13	2	146.428	225.461
12	13	3	147.803	231.485

The estimated ARMA model for the time series of AQI recorded in Velachery monitoring station is
$y_t = 4.07 + 2.54 * y_{t-1} - 2.19 * y_{t-2} + 0.65 * y_{t-3} - 1.78\varepsilon_{t-1} + 0.83\varepsilon_{t-2}$

Value of the Ljung – Box test statistic for *lag* of 20 is calculated as 14.47 with *p* value of 0.80. It indicates that there is no significant evidence to determine that the residuals are autocorrelated at a *lag* of 20.

Table 3: ARMA (3,2) Model for AQI of Velachery

p-value of Dickey-Fuller Test	Fitted Model	α	β_1, β_2 and β_3 Co-efficients	ϕ_1, ϕ_2 and ϕ_3 Co-efficients
0.0019	ARMA (3,2)	4.07	2.54, -2.19 and 0.65	-1.78 and 0.83

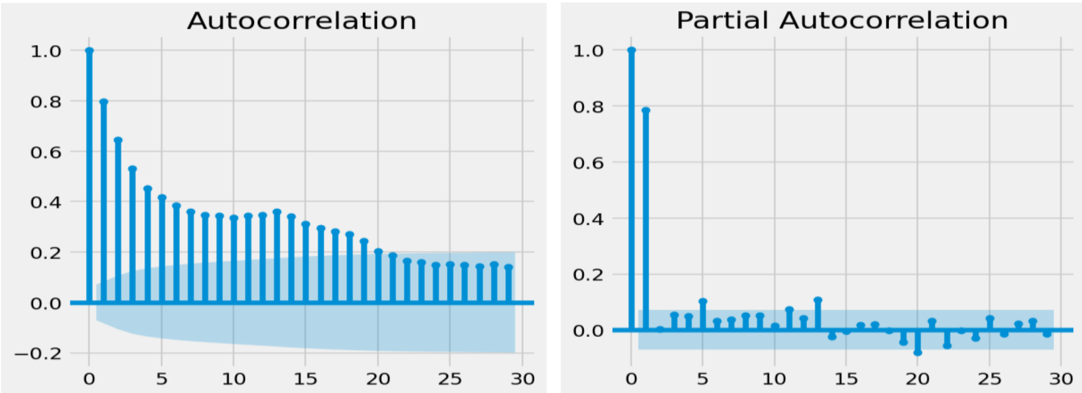

Fig. 2: ACF and PACF plots of AQI of Velachery station

4.1.2 Prophet Model for Velachery

The Prophet model is fitted by considering a linear trend growth and taking into account of forty-eight days from Jun 2022 to December 2022 declared as public holidays by the Tamil Nadu Government. The maximum change point is set at 0.09. The overall trend in the values of AQI of Velachery station during the study period is displayed in Figure 3. The plot also shows the yearly trend, holiday, weekly and yearly seasonality in the values of AQI. Pattern of trend in the values of the AQI shows a decreasing tendency over the period mentioned above. Weekly trend is almost zero in all the days, except Tuesday. The weekly trend increases from Sundays and reaches a peak on Tuesdays, which gradually falls from Wednesdays. Thursdays and Fridays show a shallow increase from Wednesdays and rise higher on Saturdays. Analysis of the yearly trend shows the highest levels of AQI during the end of June and the lowest levels during October. Spikes can be noted on holidays, which point out the influence in the values during the respective holidays.

The estimated trend function of the Prophet model for the AQI of Velachery station is

$$g(t) = \left[\left(-0.14 + \sum_{i \geq t} \delta_i\right) + 0.40 + \sum_{i \geq t} (-change\ point_i * \delta_i)\right] * y_{scaled}$$

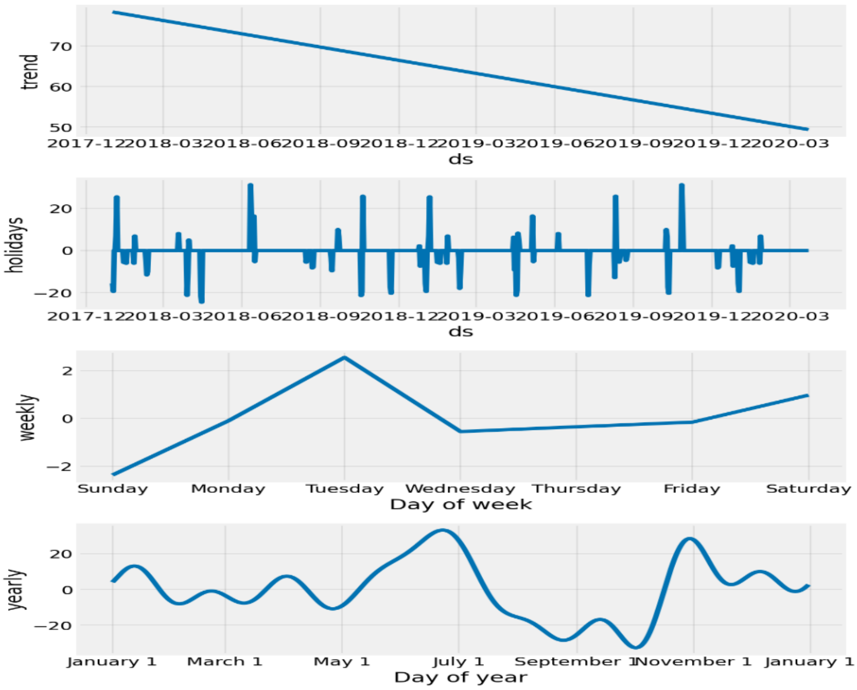

Fig. 3: Yearly, Weekly trends and Holiday effects fitted for Velachery

4.1.3LSTM Model for Velachery

Determination of the LSTM model for the time series of AQI of Velachery station shall be treated as a supervised learning problem applying the Min - Max scaling to reduce the influence of outliers. As there is no definite method for determining the appropriate number of neurons in a neural network, in this study, the model, with five hidden layers containing 50 neurons each is employed. The batch size, or the number of inputs processed at one time, is set to 100, and the model is run for 100 *epochs*. Thus, orders of the parameters of the *input gate*, *forget gate*, *cell state* and *output gate* of the LSTM network are

$W_{i,f,c,o}$ is a vector of order 1×50,

$U_{i,f,c,o}$ is a matrix of order 50× 50,

$b_{i,f,c,o}$ is a vector of order 1×50.

Here, the subscripts *i*, *f*, *c* and *o* represent respectively *input gate*, *forget gate*, *cell state* and *output gate*. Also, the weights and biases are used to control the flow of information into the LSTM unit and are updated during the training process.

4.1.4Evaluation of Time Series Models for AQI of Velachery Station

Values of the performance evaluation measures, MAE, MAPE and RMSE are computed for the estimated ARMA, Prophet and LSTM models based on the *test data* of AQI of Velachery station. The values are presented in Table 4 under the column head "Before COVID 19 Lockdown". Values of the evaluation measures show that among the three models, the LSTM model has the lowest error rate with MAE of 10.26 and RMSE of 13.05. Difference between the error rates of Prophet and ARMA models is very marginal.

The MAE and RMSE of the Prophet model are respectively 13.30 and 17.11, whereas the MAE and the RMSE of the estimated ARMA model are respectively 13.88 and 18.81. These values indicate that among the three models, accuracy level in the forecast values obtained from the estimated LSTM is relatively more. Also, there could be a margin of difference in the accuracy level of the predicted values given by the ARMA and Prophet models.

MAPE indicates that, the LSTM model achieves the highest accuracy, with the lowest error of 0.24. In comparison, the ARMA model exhibits a MAPE of 0.41, representing a 70.83% increase relative to the LSTM model, while the Prophet model shows a MAPE of 0.31. The ARMA model's error is 32.26% higher than that of the Prophet model and 70.83% higher than that of the LSTM model. These findings demonstrate that the LSTM model outperforms both the ARMA and Prophet models, establishing it as the most reliable approach, whereas the ARMA model exhibits the highest error among the three.

Table 4: MAE and RMSE Values of Time Series Models for Velachery Station

	Before COVID19 Lockdown			Post COVID19 Lockdown		
Performance Measure	ARMA	Prophet	LSTM	ARMA	Prophet	LSTM
MAE	13.88	13.30	10.26	19.01	18.67	13.52
RMSE	18.81	17.11	13.05	24.46	23.97	16.92
MAPE	0.41	0.31	0.24	0.25	0.25	0.21

The observed and the estimated values of the *test data* of AQI of Velachery station for “Before COVID19 Lockdown” period *i.e*., from 12-02-2020 to 23-03-2020 are plotted in Figure 4. The observed values are depicted by solid blue line; the estimated AQI values from the ARMA model by dashed red line; the estimated AQI values from the Prophet model by dash-dot -dash green line; and the estimated AQI values from the LSTM model by dotted magenta line. It can be noticed from Figure 4, the estimated values from the ARMA and Prophet models deviate much from the respective observed values. The estimated values from the LSTM model are relatively close to the observed values.

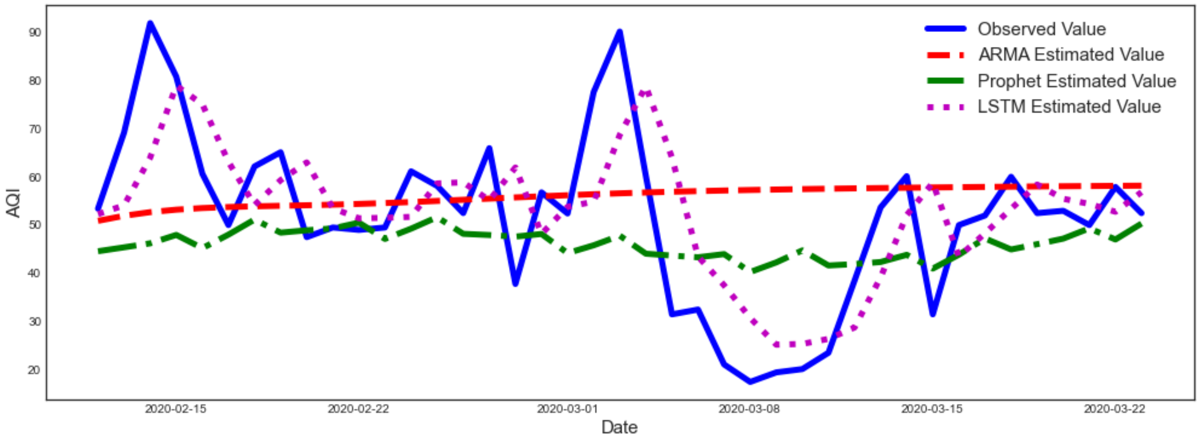

Fig. 4: Observed and Estimated Values of AQI from ARMA, Prophet and LSTM Models for Velachery Station (Before COVID19 Lockdown)

Similar analysis is carried out for “After COVID19 Lockdown” period *training data* from June 1, 2022 to November 07-2022 and *test data* from November 08, 2022 to December 05, 2022. Values of MAE and RMSE calculated from the estimated ARMA, Prophet and LSTM models are presented in Table 4 under the column head “Post COVID19 Lockdown”. Tendency in the error rates of the three models is similar to that of “Before COVID19 Lockdown” period. The MAE and RMSE for the LSTM model are 13.52 and 16.92, respectively, which are smaller than the corresponding MAE and RMSE for the ARMA and Prophet models. The observed and the estimated values plotted in Figure 5 also exhibit this fact.

The LSTM model continues to demonstrate the highest accuracy, with the lowest MAPE of 0.21. In comparison, both the ARMA and Prophet models exhibit a MAPE of 0.25, which represents a 19.05% increase in error relative to the LSTM model. It can be inferred, from all these facts, that the LSTM method of constructing time series model can provide relatively more appropriate time series model for the AQI of Velachery monitoring station.

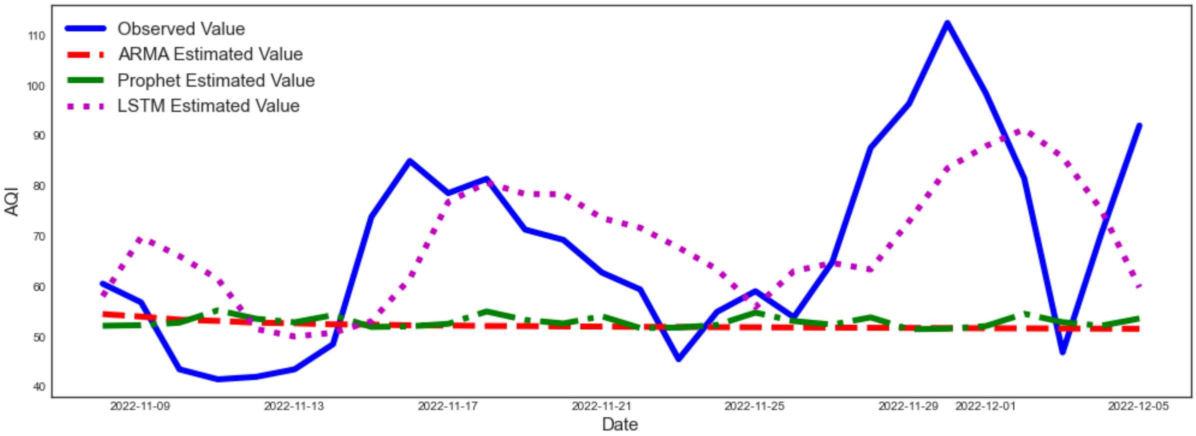

Fig. 5: Observed and Estimated Values of AQI from ARMA, Prophet and LSTM Models for Velachery Station (Post COVID19 Lockdown)

4.2 Time Series Modelling for AQI of Alandur

4.2.1ARMA Model for Alandur

The stationarity of the time series of AQI recorded in Alandur monitoring station is tested applying the Dickey-Fuller test. The *p*-value of the test statistic is computed as 0.0, which indicates that the time series of AQI recorded in Alandur monitoring station can be stationary.

The order of the ARMA model for this time series can be diagnosed from the ACF and PACF plots displayed in Figure 6. The ACF plot indicated that all *lags* are significant and the PACF plot indicate that *lags* 1, 2, 3, and 5 are significant. The ARMA models are estimated corresponding to each pair (p, q) with p = (1, 2, 3, 5) and q = (1, 2, 3). The values of AIC and BIC are calculated for each ARMA model and are displayed in Table 6. It can be noted from Table 5 that the smallest values of AIC 905.663 and BIC 935.784, is obtained corresponding to the ARMA model (3, 1) are presented Table 6. The estimated ARMA model for the time series of AQI recorded in Alandur monitoring station is

$$y_t = 4.34 + 1.32 * y_{t-1} - 0.24 * y_{t-2} - 0.09 y_{t-3} - 0.92 \varepsilon_{t-1}$$

Value of the Ljung – Box test statistic for *lag* of 20 is calculated as 14.48 with p value of 0.81. It indicates that there is no significant evidence to determine that the residuals are autocorrelated at a *lag* of 20.

Table 5: AIC and BIC Values for ARMA Models of Alandur Station

S. No	*p*	*q*	AIC	BIC
1	1	1	921.91	941.998
2	1	2	916.187	941.28
3	1	3	912.845	942.966
4	2	1	909.872	934.973
5	2	2	911.572	936.975
6	2	3	909.872	934.973
7	3	1	905.663	935.784
8	3	2	906.764	937.642
9	3	3	906.985	936.224

Table 6: ARMA (3, 1) Model for AQI of Alandur

ADF test *p* value	Fitted Model	Constant	$\beta -$ Coefficient	$\Phi -$Coefficient

0.0	ARMA (3,1)	4.34	1.32, -0.24 and -0.09	-0.92

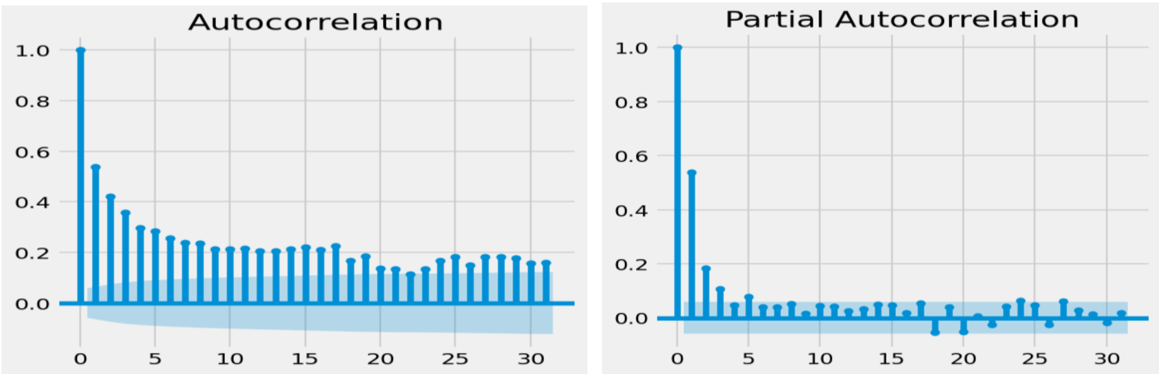

Fig. 6: Alandur ACF and PACF plot of AQI of Alandur station

4.2.2FB Prophet Model for Alandur

The Prophet model is fitted by considering a linear trend and taking into account of seventy-four days from January 2017 to March 2020 declared as public holidays by the Tamil Nadu Government. The maximum change point was set at 0.09. the overall trend in the values of AQI of Alandur station during the study period is displayed in Figure 7.

The plot also shows the yearly tend, holiday, weekly and yearly seasonality in the values of AQI. The pattern of trend in the values of AQI shows a decreasing tendency over the period mentioned above. The spikes can be point out the influence in the values during the respective holidays. The weekly trend increases from Thursdays and reaches peak on Friday, which gradually falls from Saturdays. Analysis of the yearly seasonality indicates the highest level of AQI during the beginning of January and end of October, and the lowest levels during March.

The estimated trend component of the Prophet model for the AQI of Alandur station is

$$Trend\ g(t) = \left[\left(-0.095 + \sum_{i \geq t} \delta_i\right) + 0.206 + \sum_{i \geq t} (-change\ point_i * \delta_i)\right] * y_{scaled}$$

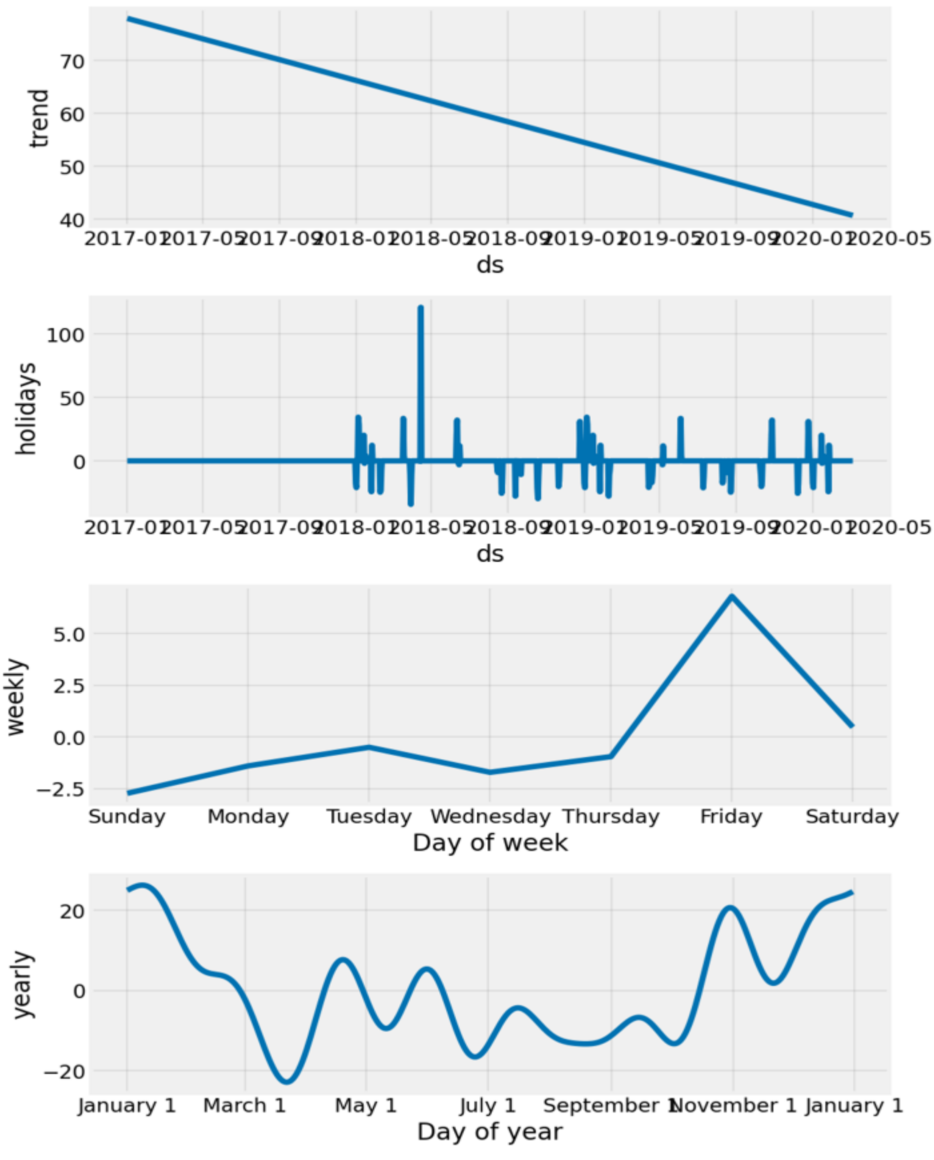

Figure 7 yearly, monthly, weekly trends and holiday plot for Alandur

4.2.3LSTM model for Alandur

Determination of the LSTM model for the time series of AQI of Alandur station shall be treated as a supervised learning problem applying the Min - Max scaling to reduce the influence of outliers. The model with five hidden layers containing 70 neurons each is employed. The batch size, or the number of inputs processed at one time, is set to 100, and the model is run for 100 *epochs*. Thus, orders of the parameters of the *input gate*, *forget gate*, *cell state* and *output gate* of the LSTM network are

$W_{i,f,c,o}$ is a vector of order 1×70,

$U_{i,f,c,o}$ is a matrix of order 70× 70,

$b_{i,f,c,o}$ is a vector of order 1×70.

Here, the subscripts *i*, *f*, *c* and *o* represent respectively *input gate*, *forget gate*, *cell state* and *output gate*. Also, the weights and biases are used to control the flow of information into the LSTM unit and are updated during the training process.

4.2.4Model evaluation for Alandur Time Series

Values of the performance evaluation measures, MAE and RMSE are computed for the estimated ARMA, Prophet and LSTM models based on the *test data* of AQI of Alandur station are presented in Table 7. The values are presented in Table 7 under the column head “Before COVID 19 Lockdown”. Vales of the evaluation measures show that among the three models, the LSTM model has the lowest error rate with MAE of 12.58 and RMSE of 17.58. The ARMA model has the highest error rate with MAE of 16.23 and RMSE of 22.90. The MAE and RMSE of estimated Prophet model are smaller than the respective values of the estimated ARMA model. The MAE and RMSE of the Prophet model are respectively 13.60 and 18.17, which is lower than the ARMA model, but higher than the LSTM model. Difference between the error rates of LSTM and ARMA model is relatively high. The LSTM and Prophet models demonstrated comparable accuracy, both achieving the lowest MAPE values of 0.25. In contrast, the ARMA model exhibited the highest error, with a MAPE of 0.36, which is 44% higher than that of the LSTM and Prophet models.

These values indicate that among the three models, accuracy level in the forecast values obtained from the estimated LSTM is relatively more. Also, there could be a margin of difference in the accuracy level of the predicted values given by the LSTM and Prophet models.

The observed and the estimated values of the *test data* of AQI of Alandur station for “Before COVID19 Lockdown” period *i.e.*, from 24-01-2020 to 23-03-2020 are plotted in Figure 8. The observed values are depicted by solid blue line; the estimated AQI values from the ARMA model by dashed red line; the estimated AQI values from the Prophet model by dash-dot -dash green line; and the estimated AQI values from the LSTM model by dotted magenta line. It can be noticed from Figure 8, the estimated values from the ARMA and Prophet models deviate much from the respective observed values. The estimated values from the LSTM model are relatively close to the observed values.

Table 7: MAE and RMSE for Alandur Time Series

		Before COVID 19 Lockdown			Post COVID 19 Lockdown		
Station/ Model	Error Measures	ARMA	Prophet	LSTM	ARMA	Prophet	LSTM
Alandur	MAE	16.23	13.60	12.58	30.72	33.79	27.66
	RMSE	22.90	18.17	17.58	37.84	38.59	37.14
	MAPE	0.36	0.25	0.25	0.32	0.37	0.26

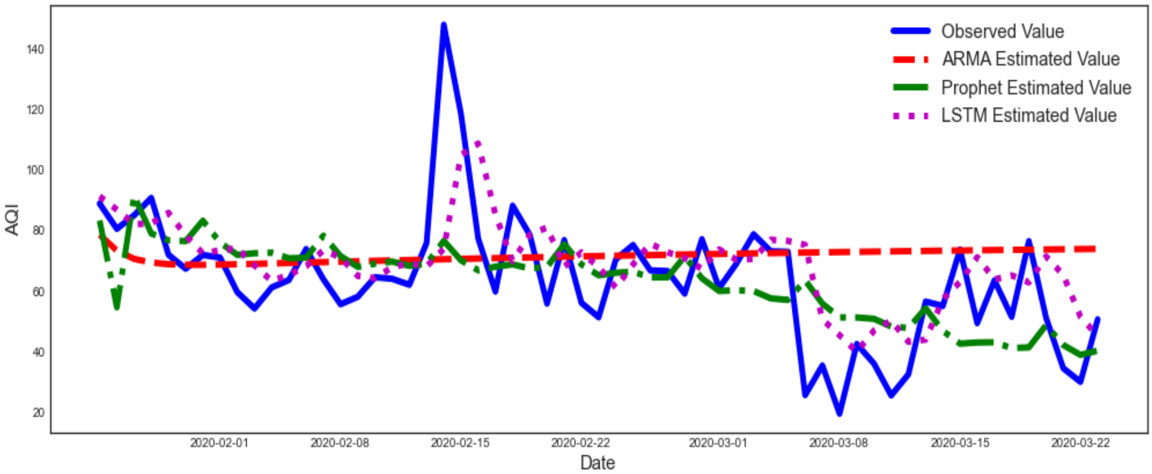

Fig. 8: Observed and Estimated values of AQI from ARMA, Prophet and LSTM Models for Alandur station

(Before COVID 19 Lockdown)

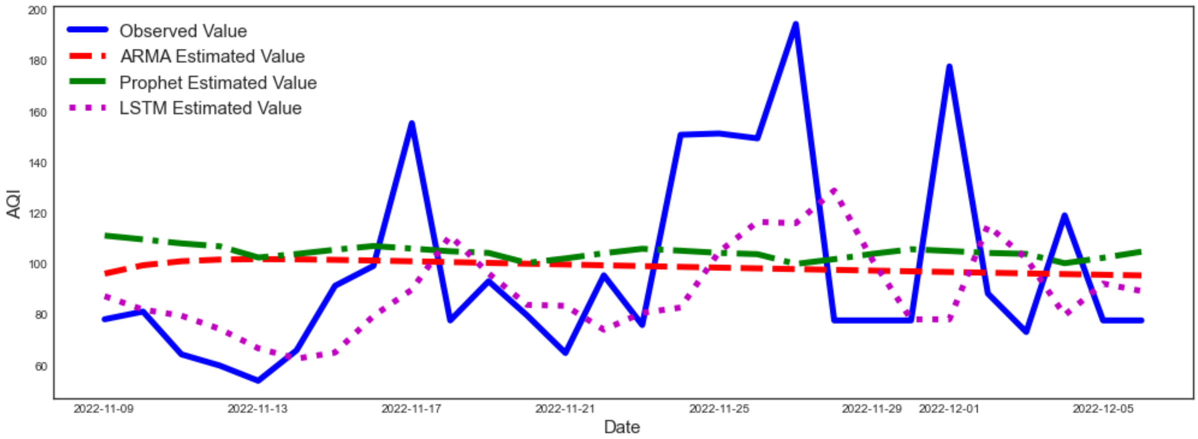

Fig. 9: Observed and Estimated values of AQI from ARMA, Prophet and LSTM Models for Alandur station (Post COVID 19 Lockdown)

Similar analysis is carried out for "After COVID19 Lockdown" period with *training data* from June 1, 2022 to November 07-2022 and *test data* from November 08, 2022 to December 05, 2022.

Values of MAE and RMSE calculated from the estimated ARMA, Prophet and LSTM models are presented in Table 7 under the column head "Post COVID19 Lockdown". Tendency in the error rates of the three models is similar to that of "Before COVID19 Lockdown" period.

The MAE and RMSE for the LSTM model are 27.66 and 37.14, respectively, which are smaller than the corresponding MAE and RMSE for the ARMA and Prophet models. The observed and the estimated values plotted in Figure 9 also exhibit this fact.

The MAE and RMSE for the LSTM model are 27.66 and 37.14, respectively, which is smaller than the MAE and RMSE for the ARMA and Prophet models. The LSTM model maintained its consistency performance, achieving the lowest MAPE of 0.26. However, the Prophet model's error increased significantly to 0.37, making it the least accurate among the three models. The ARMA model showed a slight improvement, with its MAPE decreasing to 0.32, but it still exhibited a 23.08% higher error compared to the LSTM model.

It can be inferred, from all these facts, that the LSTM method of constructing time series model can provide relatively more appropriate time series model for the AQI of Alandur monitoring station.

4.3 Time Series Modelling for AQI of Manali

4.3.1ARMA Model for Manali

The stationarity of the time series of AQI recorded in Velachery monitoring station is tested applying the Dickey-Fuller test. The *p*-value of the test statistic is computed as 0.0003, which is indicates that the time series of AQI recorded in Manali monitoring station can be stationary. The order of the ARMA model for this time series can be diagnosed from the ACF and PACF plot displayed in Figure 10.

The ACF plot indicated that *lags* 1 and 2 are significant and the PACF plot showed that *lags* 1 and 2 are significant. The ARMA models are estimated corresponding to each pair (p, q) with $p = (1, 2)$ and $q = (1, 2)$. The values of AIC and BIC are calculated for each ARMA model and are displayed in Table 8.

It can be noted from the Table 8 that the smallest value of AIC and BIC is obtained corresponding to the ARMA (1, 2) are presented in Table 9, which are respectively 528.38 and 544.86.

The estimated ARMA model for the time series of AQI recorded in Manali monitoring station is

$$y_t = 0.18 + 0.99 * y_{t-1} - 0.49\varepsilon_{t-1} - 0.21\ \varepsilon_{t-2}$$

Value of the Ljung – Box test statistic for *lag* of 20 is calculated as 23.842 with *p* value of 0.24. It indicates that there is no significant evidence to determine that the residuals are autocorrelated at a *lag* of 20.

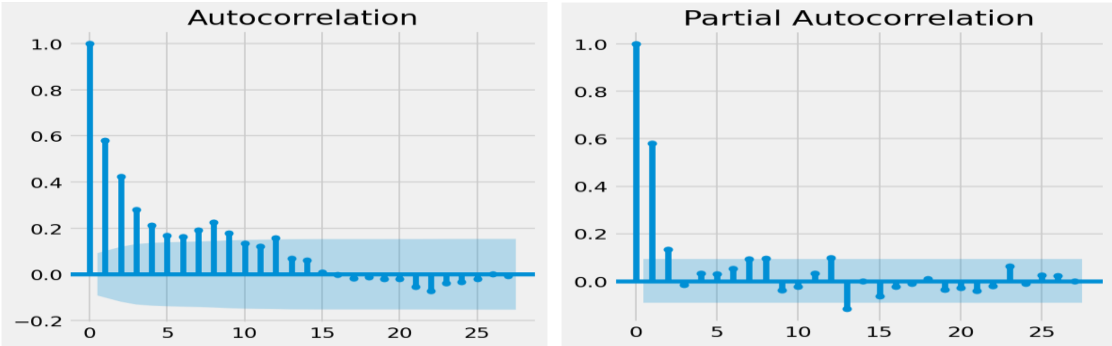

Fig. 10: ACF and PACF plot for Manali

Table 8: AIC and BIC Values for ARMA Model of Manali Station

S. No	*p*	*q*	AIC	BIC
1	1	1	544.892	557.253
2	1	2	528.378	544.860
4	2	1	507.393	507.393
5	2	2	503.755	524.356

Table 9: ARMA (1, 2) Model for AQI of Manali

Time Series	ADF test *p*-value	Fitted Model	Constant	β- Coefficient	Φ –Coefficient
Manali	0.0003	ARMA (1,2)	0.18	0.99	-0.49 and -0.22

4.3.2FB Prophet for Manali

The Prophet model is fitted by considering a linear trend growth and taking into account twenty-eight days from December 2018 to April 2020 declared as public holidays by the Tamil Nadu Government. The maximum change point is set at 0.09. The overall trend in the values of AQI of Manali station during the study period is displayed in Figure 11. The plot also shows the yearly trend, weekly, holiday and yearly seasonality in the values of AQI.

The yearly trend indicates, pattern of trend in the values of the AQI shows a decreasing tendency over the period of mentioned above. The weekly trend showed that AQI values increases from Sunday and reaches a peak on Thursday. The yearly seasonality revealed that the highest levels of AQI during the first half of October and November, and the lowest levels of AQI during the end of March and April. Spikes can be noted on holidays, which indicate the influence in the values during the respective holidays.

The estimated trend function of the Prophet model for the AQI of Manali station is

$$Trend\ g(t) = \left[\left(-0.14186748 + \sum_{i \geq t} \delta_i\right) + 0.40382178 + \sum_{i \geq t} (-change\ point_i * \delta_i)\right] * y_{scaled}$$

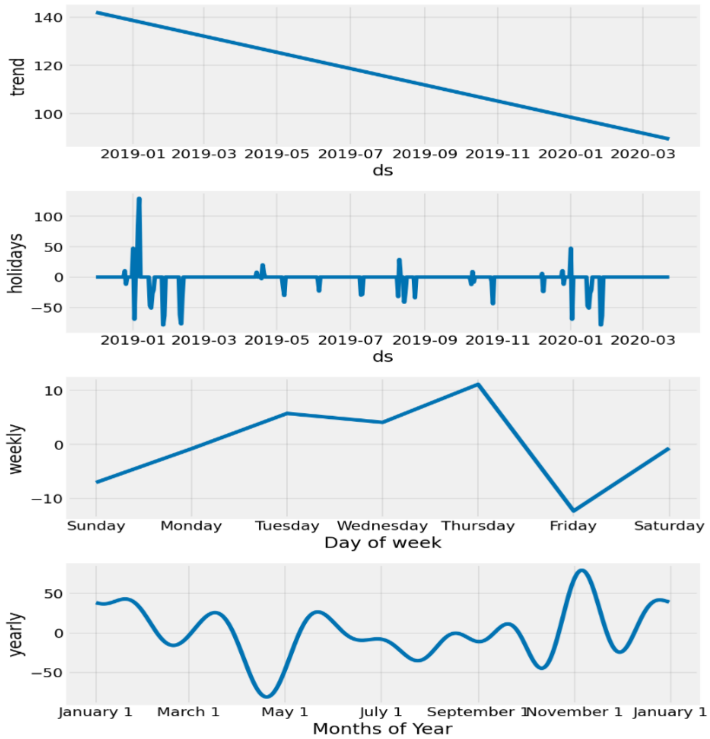

Fig. 11: Yearly, weekly trends and holiday effects fitted for Manali

4.3.3LSTM Model for Manali

This study, the model is employed multiple hidden layers, with five hidden layers containing 50 neurons each. The batch size, or the number of inputs processed at one time, is set to 50, and the model is run for 63 *epochs*. Thus, orders of the parameters of the *input gate*, *forget gate*, *cell state* and *output gate* of the LSTM network are

$W_{i,f,c,o}$ is a vector of order 1×50,

$U_{i,f,c,o}$ is a matrix of order 50× 50,

$b_{i,f,c,o}$ is a vector of order 1×50.

Here, the subscripts *i*, *f*, *c* and *o* represent respectively *input gate*, *forget gate*, *cell state* and *output gate*. Also, the weights and biases are used to control the flow of information into the LSTM unit and are updated during the training process. The *input gate*, *forget gate*, *cell state* and *output gate* weights are listed below

4.3.4Evaluation of Time Series Model for AQI of Manali Station

Values of the performance evaluation measures, MAE and RMSE are computed for the estimated ARMA, Prophet and LSTM models based on the *test data* of AQI of Manali station. The values are presented in Table 10 under the column head "Before COVID 19 Lockdown".

Values of the evaluation measures show that among the three models, the LSTM model has the lowest error rate with MAE of 11.75 and RMSE of 13.68. Difference between the error rates of Prophet and ARMA models is marginal. The MAE and RMSE of the Prophet model are respectively 24.83 and 28.62, whereas the MAE and the RMSE of the estimated ARMA model are respectively 21.79 and 25.25. These values indicate that among the three models, accuracy level in the forecast values obtained from the estimated LSTM is relatively more. Also, there could be a margin of difference in the accuracy level of the predicted values given by the ARMA and Prophet models.

The LSTM model has the lowest error rate with MAPE of 0.15. In contrast, the Prophet model had the highest error, with a MAPE of 0.33, followed by the ARMA model with a MAPE of 0.30.These findings indicate that while the LSTM model was the most reliable before the COVID-19 lockdown period.

Table 10: MAE and RMSE Values of Time Series for Manali Station

		Before COVID 19 Lockdown			Post COVID 19 Lockdown		
Station/ Model	Error Measures	ARMA	Prophet	LSTM	ARMA	Prophet	LSTM
Manali	MAE	21.79	24.83	11.75	17.33	15.09	14.95
	RMSE	25.25	28.62	13.68	24.25	21.53	21.65
	MAPE	0.30	0.33	0.15	0.18	0.15	0.16

The observed and the estimated values of the *test data* of AQI of Manali station for "Before COVID19 Lockdown" period *i.e.*, from 29-02-2020 to 23-03-2020 are plotted in Figure 12. The observed values are depicted by solid blue line; the estimated AQI values from the ARMA model by dashed red line; the estimated AQI values from the Prophet model by–dash-dot -dash green line; and the estimated AQI values from the LSTM model by dotted magenta line. It can be noticed from Figure 12, the estimated values from the ARMA and Prophet models much from the respective observed values. The estimated values from the LSTM model are relatively close to the observed values.

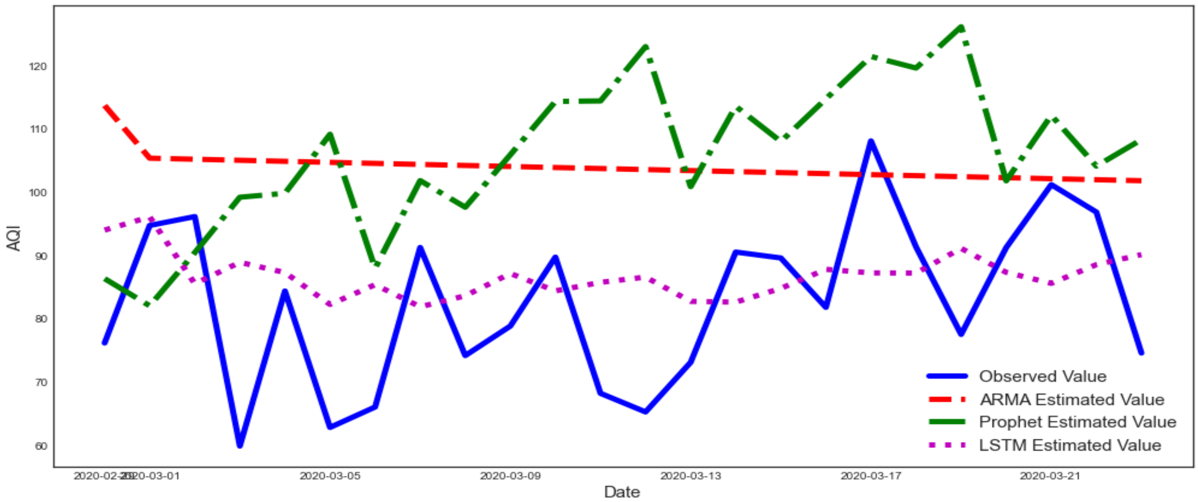

Fig.12: Observed and Estimated Values of AQI from ARMA, Prophet and LSTM Models for Velachery Station (Before COVID19 Lockdown)

Similar analysis is carried out for "After COVID 19 Lockdown" period with *training data* from June 1,2022 to November 07-2022 and *test data* from November 08, 2022 to December 05, 2022. Values of MAE and RMSE calculated from the estimated ARMA, Prophet and LSTM models are presented in Table 10 under the column head "Post COVID19 Lockdown". Tendency in the error rates of the three models is similar to that of "Before COVID19 Lockdown" period.

The MAE and RMSE for the LSTM model are 17.33 and 24.25, respectively, which is smaller than the corresponding MAE and RMSE for the ARMA and Prophet models. The observed and the estimated values plotted in Figure 13 also exhibit this fact. The difference between the error rates of LSTM and Prophet model is marginal. In the post-lockdown period, all models demonstrated improved accuracy. The Prophet model achieved the lowest MAPE of 0.15, followed closely by the LSTM model with 0.16, while the ARMA model recorded a MAPE of 0.18. Compared to the pre-lockdown period, the ARMA and Prophet models showed notable improvements, while the LSTM model's error slightly increased.

It can be inferred, from all these facts, that the LSTM method of constructing time series model can provide relatively more appropriate time series model for the AQI of Manali monitoring station.

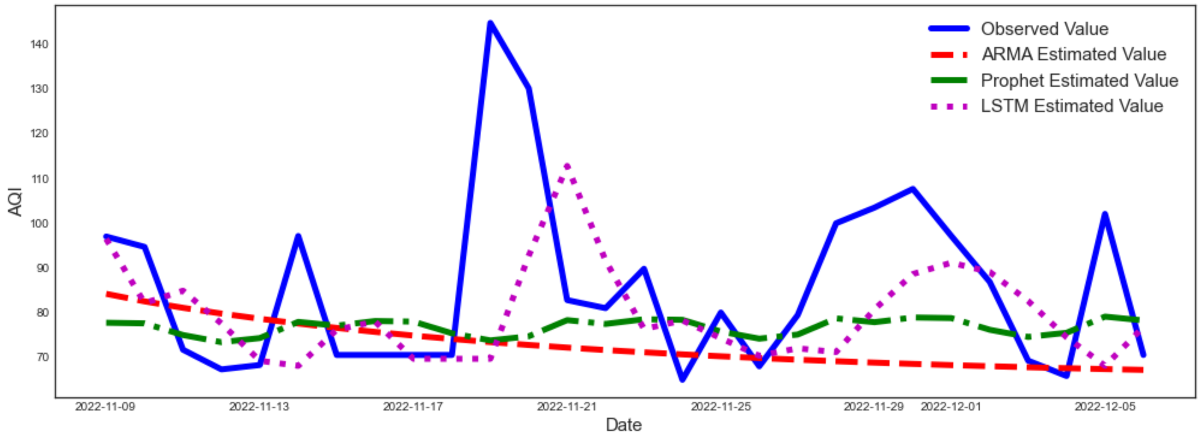

Fig. 13: Observed and Estimated Values of AQI from ARMA, Prophet and LSTM Models for Velachery station (Post COVID19 Lockdown)

5. CONCLUDING REMARKS

Time series models have special feature of forecasting the observation for future time period. They play vital role in many fields, particularly in forecasting economy, demography and environment of a country. Several methods have been developed with different approaches and performance of the methods are found to vary with respect to the data.

An empirical study is carried out in this work to compare the performance of the ARMA method developed in the statistical approach, the Prophet method developed employing Machine Learning procedure and the long short-term memory method developed employing Deep Learning methodology. Time series models are constructed for daily air quality indices observed during before and after COVID 19 lockdown periods of three locations of Chennai city. Among the three methods, the long short-term memory method performs relatively better than the other two methods for the three time series with respect to measures of accuracy in forecast values. Performance of the ARMA method and the Prophet method depend upon the data set. Results do not differ, except numerical values, for pre and post COVID 19 lockdown periods.

Author Contributions: For research articles with multiple authors, include a brief paragraph outlining each author's contributions using the following format: "Conceptualization, Sumithra Palra. and Loganathan Appaia.; methodology, Sumithra Palra. and Loganathan Appaia ; software, Sumithra Palraj.; validation, Loganathan Appaia., Deneshkumar Venugopal. and Gunasekaran Munian.; formal analysis, Sumithra Palraj.and Loganathan Appaia; investigation, Loga-nathan Appaia and Deneshkumar Venugopal; resources, Sumithra Palraj.; data curation, Sumithra Palraj and Loga-nathan Appaia.; writing—original draft preparation, Sumithra Palraj and Loganathan Appaia.; writing—review and ed-iting, , Loganathan Appaia., Deneshkumar Venugopal. and Gunasekaran Munian.; visualization, Sumithra Palraj.; su-pervision, , Loganathan Appaia., Deneshkumar Venugopal. and Gunasekaran Munian.; project administration, Loga-nathan Appaia.; funding acquisition, Loganathan Appaia. All authors have read and agreed to the published version of the manuscript." Authorship should be restricted to individuals who have made significant contributions to the re-search.

Funding: "This research received no external funding."

Informed Consent Statement: "Not applicable." (for studies not involving humans)

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections

Conflicts of Interest: Authors must declare any conflicts of interest or state, "The authors declare no conflicts of interest." It is essential to identify and disclose any personal circumstances or interests that may be perceived as influencing the representation or interpretation of the reported research results. Additionally, any involvement of funders in the design of the study, data collection, analysis, interpretation, manuscript writing, or decision to publish must be clearly stated. If there was no role for the funders, please include the statement: "The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results."

REFERENCES

Abhilash, M.S.K., Thakur, A., Gupta, D. and Sreevidya, B. (2018) 'Time series analysis of air pollution in Bengaluru using ARIMA model', in *Ambient Communications and Computer Systems*. Springer, pp. 413–426.

Ammasi Krishnan, M., Devaraj, T., Velayutham, K., Perumal, V. and Subramanian, S. (2020) 'Statistical evaluation of PM2. 5 and dissemination of PM2. 5, SO2 and NO2 during Diwali at Chennai, India', *Natural Hazards*, 103(3), pp. 3847–3861.

Anand Kumar Varma S, Mahmood Anas M S, Harun Raseed M, Nithishbalasubramanian O and Madhan Kumar R (2021) 'Determination of air quality index and its impacts on human health in Chennai City', *Magna Scientia Advanced Research and Reviews*, 3(1). Available at: https://doi.org/10.30574/msarr.2021.3.1.0069.

Angelena, J.P., Stanley Raj, A., Viswanath, J. and Muthuraj, D. (2021) 'Evaluation and forecasting of PM10 air pollution in Chennai district using Wavelets, ARIMA, and Neural Networks algorithms', *Pollution*, 7(1), pp. 55–72.

Box, G.E.P., Jenkins, G.M. and Reinsel, G. (1970) 'Time series analysis: forecasting and control Holden-day San Francisco', *BoxTime Series Analysis: Forecasting and Control Holden Day1970* [Preprint].

Briggs, W.M., Makridakis, S., Wheelwright, S.C., Hyndman, R.J. and Diebold, F.X. (1999) *Forecasting: Methods and Applications*, *Journal of the American Statistical Association*. John wiley & sons. Available at: https://doi.org/10.2307/2669717.

Büyükşahin, Ü.Ç. and Ertekin, Ş. (2019) 'Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition', *Neurocomputing*, 361, pp. 151–163. Available at: https://doi.org/10.1016/J.NEUCOM.2019.05.099.

Calvo, B. and Santafé Rodrigo, G. (2016) 'scmamp: Statistical comparison of multiple algorithms in multiple problems', *The R Journal, Vol. 8/1, Aug. 2016* [Preprint].

Gunasekar, S., Joselin Retna Kumar, G. and Dileep Kumar, Y. (2022) 'Sustainable optimized LSTM-based intelligent system for air quality prediction in Chennai', *Acta Geophysica* [Preprint]. Available at: https://doi.org/10.1007/s11600-022-00796-6.

Haviluddin, Alfred, R., Obit, J.H., Hijazi, M.H.A. and Ibrahim, A.A.A. (2015) 'A performance comparison of statistical and machine learning techniques in learning time series data', *Advanced Science Letters*, 21(10). Available at: https://doi.org/10.1166/asl.2015.6490.

Janarthanan, R., Partheeban, P., Somasundaram, K. and Navin Elamparithi, P. (2021) 'A deep learning approach for prediction of air quality index in a metropolitan city', *Sustainable Cities and Society*, 67. Available at: https://doi.org/10.1016/j.scs.2021.102720.

Kihoro, J.M., Athiany, H., WalterO, Y. and W, K.H. (2013) 'Imputation of incomplete non-stationary seasonal time series data', 3(12). Available at: www.iiste.org.

Laxmipriya, S. and Narayanan, R.M. (2021) 'COVID-19 and its relationship to particulate matter pollution–Case study from part of greater Chennai, India', *Materials Today: Proceedings*, 43, pp. 1634–1639.

Mani, G. and Volety, R. (2021) 'A comparative analysis of LSTM and ARIMA for enhanced real-time air pollutant levels forecasting using sensor fusion with ground station data', *Cogent Engineering*, 8(1). Available at: https://doi.org/10.1080/23311916.2021.1936886.

Nadeem, I., Ilyas, A.M. and Sheik Uduman, P.S. (2020) 'Analyzing and forecasting ambient air quality of chennai city in india', *Geography, Environment, Sustainability*, 13(3), pp. 13–21. Available at: https://doi.org/10.24057/2071-9388-2019-97.

Naveen, V. and Anu, N. (2017) 'Time series analysis to forecast air quality indices in Thiruvananthapuram District, Kerala, India', *J Eng Res Appl*, 7(6), pp. 66–84.

Parmezan, A.R.S., Souza, V.M.A. and Batista, G.E.A.P.A. (2019) 'Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model', *Information Sciences*, 484. Available at: https://doi.org/10.1016/j.ins.2019.01.076.

Patra, S.R. (2017) 'Time series forecasting of air pollutant concentration levels using machine learning', *Adv Comput Sci Inf Technol (ACSIT)*, 4(5), pp. 280–284.

Samal, K.K.R., Babu, K.S., Das, S.K. and Acharaya, A. (2019) 'Time series based air pollution forecasting using SARIMA and prophet model', in *ACM International Conference Proceeding Series*. Available at: https://doi.org/10.1145/3355402.3355417.

Samia, A., Kaouther, N. and Abdelwahed, T. (2012) 'A hybrid ARIMA and artificial neural networks model to forecast air quality in urban areas: case of Tunisia', in *Advanced Materials Research*. Trans Tech Publ, pp. 2969–2979.

Sánchez-Pozo, N.N., Trilles-Oliver, S., Solé-Ribalta, A., Lorente-Leyva, L.L., Mayorca-Torres, D. and Peluffo-Ordóñez, D.H. (2021) 'Algorithms Air Quality Estimation: A Comparative Study of Stochastic and Heuristic Predictive Models', in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, pp. 293–304.

Sekar, T., Venkatachalam,A.V.V., Kallolikar, S. (2020) *Impact of Covid-19 Lockdown on Air Quality of Chennai*.

Taylor, S.J. and Letham, B. (2018) 'Forecasting at Scale', *American Statistician*, 72(1). Available at: https://doi.org/10.1080/00031305.2017.1380080.

Toharudin, T., Pontoh, R.S., Caraka, R.E., Zahroh, S., Lee, Y. and Chen, R.C. (2020) 'Employing long short-term memory and Facebook prophet model in air temperature forecasting', *Communications in Statistics: Simulation and Computation* [Preprint]. Available at: https://doi.org/10.1080/03610918.2020.1854302.

Zhang, X., Liu, Y., Yang, M., Zhang, T., Young, A.A. and Li, X. (2013) 'Comparative study of four time series methods in forecasting typhoid fever incidence in China', *PloS one*, 8(5), p. e63116.