The University of Tehran Press

Pollution

Print ISSN: 2383-451X
Online ISSN: 2383-4501

https://jpoll.ut.ac.ir/

Forecasting the Short-Term Changes of Surface Ozone and NO_2 during a Festival Event Using Stochastic and Neural Network Models

Ebin Antony[1] | Keerthi Lakshmi[2] | Sunil Kumar[1] | Nishanth Theeyancheri[2]✉ | Jalaja Kunnath[3] | Satheesh Kumar[4] | Annie Sabitha Paul[5]

1. Department of Information Technology, Kannur University, Kannur, India
2. Department of Physics, Sree Krishna College Guruvayur, Kerala, India- 680102
3. Department of Applied Science and Humanities, Nehru College of Engineering and Research Centre, Pampady, Thrissur, Kerala-680588
4. Department of Atomic and Molecular Physics, Manipal Academy of Higher Education, Karnataka, India- 576104
5. Department of Applied Sciences, Govt. Engineering College Kannur, Kerala India-670563

Article Info

Article type:
Research Article

Article history:
Received: 25 May 2024
Revised: 1 February 2025
Accepted: 1 June 2025

Keywords:
Ozone,
NO_2,
Fire bursting,
Forecasting,
ANN,
SARIMA

ABSTRACT

Air pollution is one of the most destructive environmental issues on the local, regional, and global level. Its negative influences go far beyond ecosystems and the economy, harming human health and environmental sustainability. By these facts, efficient and accurate modelling and forecasting the concentration of air pollutants are vital. Hence, this work explores investigate the time series components of surface ozone (O_3) and its precursor nitrogen dioxide (NO_2) and develops a model for predicting O_3 variations produced by intense fireworks during the Vishu festival over Kannur. Time series methods using Stochastic and Recurrent Neural Network (RNN) and Seasonal Autoregressive Integrated Moving Average (SARIMA) models are considered the most accurate tools for estimating air pollution trends due to their logical flexibility. Model performance is evaluated based on statistical measurements indicating an increasing trend in O_3 concentration of 0.11 ppb/year and NO_2 of 0.18 ppb/year. Based on the analysis, we found that the SARIMA model shows better accuracy with a Mean Squared Error (MSE) of 0.55 and a Root Mean Squared Error (RMSE) of 0.74. The broader implications of this study highlight the applicability of advanced time series forecasting techniques for air quality monitoring during short-term pollution events.

Cite this article: Antony, E., Lakshmi, K., Kumar, S., Theeyancheri, N., Kunnath, J., Kumar S. & Sabitha Paul, A. (2025). Forecasting the Short-Term Changes of Surface Ozone and NO_2 during a Festival Event Using Stochastic and Neural Network Models. *Pollution*, 11(3), 576-592.
https://doi.org/10.22059/poll.2025.377011.2387

DOI: https://doi.org/10.22059/poll.2025.377011.2387

INTRODUCTION

Air pollution is considered as a serious environmental threat that influences global climate and public health. The increasing concentrations of greenhouse gases and particulate matter (PM) in the atmosphere induce changes in the radiative forcing of the atmosphere which leads to poor air quality over a location (Keerthi Lakshmi et al., 2024). Surface O_3 is a secondary air pollutant and a strong greenhouse gas with a high oxidising capacity, and it plays a key role in the photochemistry of the atmosphere to change the radiative forcing at the Earth's surface

*Corresponding Author Email: *nisthu.t@gmail.com*

(Monks et al., 2015). Inhalation of O_3 rich air results in severe health issues including reduced lung function and oxidative stress (Maji and Namdeo 2021; Li et al., 2016).

Fireworks related to festival celebrations are one of the potential sources of surface O_3 produced by air pollution. An increase in surface O_3 concentration was observed in the fireworks during Vishu festival in Kerala, a coastal location in south Indian state. Vishu festival is celebrated on the zodiacal sign of the regional New Year, which usually falls in mid-April. The spectacular display of fireworks is the main attraction of this festival which starts on Vishu eve (14 April) by setting off fireworks late at night and then continues up to the early morning of Vishu day (15 April) with two distinct spells of fire bursting episodes. Only limited reports are available describing the detailed chemistry of ozone production during fireworks associated with Indian festivals (Resmi et al., 2021; Nishanth et al., 2012; Attri et al., 2001). The air quality becomes very poor during the fireworks and a forecast of the air quality well before the festival is highly useful to issue proper warnings.

Very often, air quality forecasts are performed with the help of analytical air pollution models. The primary objective of time series modelling is to produce credible models by gathering and evaluating data from the past. It is the process of using a model to predict future values based on observed historical values. Forecasting time series finds many applications in various areas of science that researchers have to work harder to make a model that matches the data. Artificial neural networks (ANN) have recently been identified as a potential tool for modeling time series and making predictions. The most extensively used ANNs in forecasting problems are multilayer perceptrons (MLPs). Since the number of input neurons helps reveal the relation between observations in MLPs, it is quite effective in the performance of the network. A feed forward network is a single hidden layer found in the majority of multilayer perceptron models.

Hamzacebi (2008) developed a Seasonal Artificial Neural Network (SANN) model for seasonal time series forecasting. This basic seasonal time series forecasting method has proven to be more effective in a system showing nonlinear behavior. Many methods have been reported in the literature; most of these are based on the use of neural networks and autoregressive integrated moving average (ARIMA) models (Capilla 2016; Paoli et al 2011; Coman et al., 2008; Basurko et al 2007; Duenas et al 2005; Kumar et al 2004). Both the ARIMA and the Box-Jenkins models are well-known for their accuracy and forecasting (Kumar et al 2004).

Even though the ARIMA model has been successful in various forecasting purposes, it suffers from restriction to its linear form. Very often, linear models are inconsistent for composite real-world problems. Thus, input hidden and output neuron numbers are quite vital for increasing the performance of ANN forecasting. Subsequently, Recurrent Neural Network (RNN) has become popular due to its repeatedly transmitted input through the loop, which results a considerable modification to the neural network model's measurements. Both are also noted for how effectively they can predict the future. For this method, the time series must be linear and have statistics that resembles the normal distribution SARIMA model was made by Box and Jenkins to predict seasonal time series data (Kumar et al 2004; Beldjillali et al 2016).

The inherent ability of Artificial Neural Networks (ANNs) to express nonlinearity without assuming the statistical distribution of data makes them ideal for time series forecasting. When the number of error gradients goes up during an update, the network becomes unstable. Bursting happens quickly when gradients in the network layer with values greater than 1 are multiplied over and over again, but it doesn't happen when the values are less than 1. Scientists had to come up with a new RNN model called long short-term memory to fix these problems (LSTM). Long short memory can tackle this problem by controlling how people remember things through gates. There are many other neural network structures in the literature because research in this field is always going on. But for this work, we will focus on the stochastic and recurrent neural network (RNN) forecasting models. Several research papers are available using ARIMA models for the forecast of ozone levels across the globe (Samuel Selvaraj et.al 2013; Arputharaj et al 2016). Generally, in atmospheric models, time series forecasting is different from numerical weather forecasting because it uses data from the past to make predictions about the future.

Hindcast is recognised as a versatile tool related to the process of using historical information or data to simulate past conditions or other naturally occurring phenomena from the past. This process involves deploying computer models or mathematical simulations, then using the collected data, to understand and investigate potential historical events. Hindcasting techniques are commonly used in climatological, marine, and other fields to confirm and improve models, explain past patterns, and predict future conditions.

The discrepancy between predicted and observed outcomes in various environmental episodes suggests positive results, and the gap between these results leads to dynamics of scientific theories. Xu Wanyun et al., (2016) and (2018) conducted a comprehensive study of surface O_3 chemistry and transport at the north-eastern Tibetan Plateau region in China from 1994 to 2013 using the Mann-Kendall test and Hilbert-Huang transform analysis to investigate the trend and periodicity of O_3. They found that the influence of valley breeze fluctuations caused unusual differences in night and day O_3 and seasonal maximum and minimum concentrations due to the exchange between the stratosphere and troposphere, revealing boundary layer dynamics. The first documented evidence of an inverse relationship between tropospheric and stratospheric ozone, due to the low NOx regime of the marine atmosphere, was observed in the marine atmosphere at the tropical site in the Eastern Pacific during a solar cycle and the observed change was significantly larger than estimated from the photochemical model (Chandra S et al., 1999). Zhang Aoxing et al. 2023, have developed an efficient 2D convolutional neural network surface ozone ensemble prediction system (2DCNN-SOEF) that showed performance comparable to current operational prediction systems and fairly good accuracy required by the Chinese authorities with a lead of up to 144 hours of fulfilled time. This makes their ensemble forecasting framework versatile to forecast other meteorologically dependent environmental risks worldwide. In Poland, multiple linear regression (MLR) and artificial neural network (ANN) models were investigated for each season separately using temperature, relative humidity, time of day and 1-day lagged surface ozone values.

The performance of ANN was slightly better than the MLR model and the statistical models showed a better performance in all seasons, except in winter (Pawlak et al, 2023). Hata Hiroo et al. 2023 quantified the emission inventory of primary air pollutants in Japan to 2050 using WRF v 4.3.1 and inputs from the Japanese government socio-economic model. Their findings suggest that the implementation of net-zero carbon technology can result in a significant reduction of primary emissions of NOx, SO2, and CO by 50-60%, as well as a decline of 10% in primary emissions of volatile organic compounds (VOCs) and PM2.5 by 30%.

Here we tried to forecast O_3 concentrations during Vishu festival using RNN, ARIMA and SARIMA and a comparison of the efficiencies of these models in forecasting the concentration of air pollutants in the atmosphere with better precision and the identification of the efficiencies appropriate model.

METHODOLOGY

Description about the study area

Kannur is the northern district of Kerala lying along the coastal belt of the Arabian Sea, which makes it one of the most popular tourist destinations in South India. Detailed description of the analyser used for the study is reported in our earlier publication (Resmi et al. 2021).

Data preprocessing and decomposition of time series

The initial stage in data science for classification and information retrieval problems is preprocessing. Before being processed by any machine learning or data mining algorithms, the raw data that has been obtained from the actual world would first go through pre-processing procedures. Data cleaning, data transformation, and data reduction are all part of the data preparation process. Missing data, as well as noise or outlier removal, are addressed during data cleaning. The outlier handled using Inter Quartile Range (IQR) and Z-score. The missing data handled using the linear interpolation and KNN imputation (Antony et al., 2021). There are two approaches that are frequently used to break down time series into their component parts, including (i) decomposition by additive hypothesis and (ii) decomposition by multiplicative hypothesis. The following time series has been represented as eq. (1) using the decomposition by additive hypothesis method:

$$y_t = \mathrm{T_t} + \mathrm{S_t} + \mathrm{C_t} + \mathrm{R_t} \tag{1}$$

Here, y_t, T_t, S_t, C_t, and R_t represent the time series, trend, seasonal, cyclic, and random fluctuations at time t respectively.

Sen's slope estimator and Box-Jenkins modeling for trend test

Magnitude of the trend may be predicted using Sens estimator (Sen., 1968). In accordance with Sen, the slope (Ti) of each and every data pair may be calculated as follows:

$$\mathrm{Ti} = \frac{x_j - x_k}{j-k} \text{ for } i = 1,2\ldots\ldots N \tag{2}$$

where x_j and x_k are the data values (j > k) at time j and time k respectively, and the slope estimate of Sen estimate is the median of these N values of Ti and is given by

$$Q_i = \begin{cases} T_{\frac{N+1}{2}} & \text{N is odd} \\ \frac{1}{2}\left(T_{\frac{N}{2}} + T_{\frac{N+2}{2}}\right) & \text{N is even} \end{cases} \tag{3}$$

When the value of Qi is positive, the time series is exhibiting an upward trend, and when the value of Qi is negative, the time series is exhibiting a downward tendency (Asfaw et al 2017).

Time series modelling and forecasting frequently employ the Box-Jenkins modelling method. The four steps of the Box-Jenkins approach are tentative identification (I), parameter estimation (II), diagnostic verification (III), and forecasting (IV). When building a Box-Jenkins model, the first thing that has to be determined is whether or not a time series is stationary and whether or not there is substantial seasonality that needs to be taken into mind (Mashfiqul Huq et al 2018). Formal tests, partial autocorrelation functions, and line chart autocorrelation functions are used to determine whether a variable is stationary. Non-stationarity in a time series can be recognized when a line plot exhibits trend, seasonality, and a very slowly decaying autocorrelation plot. This is called random walk-in econometrics, and it can detect non-stationarities even in the absence of trends.

Dickey and Fuller (1979), Phillips and Perron (1986) and Kwiatkowski et al., (1992) have provide foundation of official test for stationarity. When it comes to the identification step, order of the models is affected by both autocorrelation function and partial autocorrelation function. The greatest likelihood approach is then used to conclude the estimate phase. The Akaike (1974) information criteria and residual variance are examples of model selection criteria. In order to decide which model is superior, we put the modified Akaike information criteria (Nariaki Sugiura1978) and the Bayesian information criterion (Gideon Schwarz., 1978) to use. Standardized residual plots of outliers, ACF plots, and Ljung-Box (178) test statistics of residuals are used for diagnostic validation. These are used to check for white noise. The stability of the model that was chosen may then be examined by using the mean squared error and root mean squared error. At long last, the prediction process may begin with the best model that was picked.

ARIMA, SARIMA and RNN model

The ARMA model's capabilities are extended in the Autoregressive Integrated Moving Average ARIMA model. It was used in situations when there were indications that the data was not stationary. The letters ARIMA(p,d,q), which stand for autoregressive, integrated, and moving average, respectively, indicate the ordering of the autoregressive, integrated, and moving average components of the model. The p, d, and q in this abbreviation stand for the autoregressive, integrated, and moving average components, respectively. An equation that describes an ARIMA (p,d,q) model may be expressed as

$$\emptyset(\mathrm{B})\nabla^{\mathrm{d}}\mathrm{y}_{\mathrm{t}} = \mathrm{c} + \theta(\mathrm{B})\varepsilon_{\mathrm{t}} \text{ with } \{\varepsilon_{\mathrm{t}}\}\sim \mathrm{WN}(0,\sigma^2) \quad (4)$$

Where, *WN* stands for white noise.
$\nabla^d = (1-B)^d$ (The *d* order differencing operator)

$\emptyset(\mathrm{B}) = 1 - \emptyset_1\mathrm{B} - \emptyset_2\mathrm{B}^2 - \cdots - \emptyset_{\mathrm{p}}\mathrm{B}^{\mathrm{p}}$ The *p* order of AR operator)

$\theta(\mathrm{B}) = 1 + \theta_1\mathrm{B} + \theta_2\mathrm{B}^2 + \cdots + \theta_{\mathrm{q}}\mathrm{B}^{\mathrm{q}}$ (The *q* order of MA operator)

ε_t is random shocks, *c* is constant and y_t is any time series.
The standard ARIMA model gets renamed to the Seasonal ARIMA (SARIMA) model if it is used to a time series that exhibits a seasonal influence. The equation for a generic SARIMA model, which is denoted by the notation SARIMA (p,d,q), is as follows

$$(B^s)\emptyset(\mathrm{B})\nabla_s^D\nabla^d y_t = c + \Theta(B^s)\theta(B)\varepsilon_t \quad (5)$$

Where,
$\boldsymbol{\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}}$ (The P order of seasonal AR operator)

$\boldsymbol{\Theta(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdots + \Theta_Q B^{QS}}$ (The Q order of seasonal MA operator)

Both $\emptyset$(B) and θB are equivalent to the same thing in the equation (4). $\nabla^d = (1-B)^d$ and $\nabla_S^D = (1-B)^D$ refer to non-seasonal difference operators as well as seasonal difference operators. c is a constant, yt is a time series, and the standard Gaussian white noise process is represented by ε_t.

Recurrent Neural Network (RNN) can handle data in a certain order. We use them in this case to work with time series. Recurrent neural networks make predictions based on both the information they are given and the results they have given in the past. This concept is highly sensible; we could build neural networks that advance values in time. But simple solutions like this rarely work the way they are supposed to. It's hard to teach them things, and they forget. Instead, we need a machine with some form of memory. Long-term memory and gated recurrent unit are two common and efficient RNN models.

Long-term short-term memory (LSTM)

Hochreiter and Schmidhuber (1997) showed that long-term short-term memory is a neural network memory unit controlled by gates. It consists of three gates that control how the memory works. Backpropagation could be used to learn the weights for these simple weighted-sum logistic functions. It shows that the LSTM fits right in, even if the neural network and its training process seem complicated. Without additional training or optimization, it is capable of learning what necessary, storing relevant information in memory is, and retrieving that information when needed. Cell state (9), representing long-term memory, is controlled by the input gate (6) and forget gate (7). Priority is given to the information stored in memory location, which corresponds to the output vector or hidden state (10) created by the output gate (8). This memory mechanism provides the network with long-term recall, a skill that was notably missing in pure recurrent neural networks.

$$i_t = \mathrm{sigmoid}\left(W_i x_i + U_i h_{t-1} + b_i\right) \tag{6}$$

$$f_t = \mathrm{sigmoid}(W_f x_t + U_f h_{t-1} + b_f \tag{7}$$

$$o_t = \mathrm{sigmoid}(W_o x_t + U_o h_{t-1} + b_0) \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{9}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{10}$$

Confidence Interval (CI)

After generating predictions using a time series model, such as SARIMA or a Recurrent Neural Network (RNN), the residuals defined as the differences between the observed values and the corresponding predicted values are computed.

$$\text{Residual} = \text{Observed} - \text{Predicted} \tag{11}$$

The standard deviation (σ) of these residuals is then calculated to quantify the variability or dispersion in the model's prediction errors. This standard deviation provides a measure of how well the model has captured the underlying data patterns and is a key step in assessing the accuracy of the predictions and constructing confidence intervals for future forecasts.

For a 95% confidence interval, it is typically assumed that the residuals follow a normal distribution, allowing for the application of the Z-score from a standard normal distribution, specifically 1.96 for a 95% confidence level. This assumption is valid under the premise that the residuals are normally distributed. In cases where the residuals deviate from normality, alternative approaches, such as utilizing a different distribution or employing bootstrapping techniques, may be necessary to ensure accurate interval estimation.

The Standard Error (SE) of the predictions is calculated by dividing the standard deviation (σ) by the square root of the number of data points(n) used in the prediction:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}\mathrm{SE} = \frac{\sigma}{\sqrt{n}} \tag{12}$$

The margin of error (MOE) is calculated by multiplying the standard error (SE) by the Z-score corresponding to the desired confidence level. For a 95% confidence interval, the

Z-score is 1.96. Thus, the margin of error is determined as follows:

$$MOE = 1.96 \times SEMOE = 1.96 \times SE \tag{13}$$

This provides the range within which the true population parameter is expected to fall with 95% confidence. The 95% confidence interval is computed by adding and subtracting the margin of error (MOE) from the predicted value. Mathematically, this is expressed as:

$$CI = Predicted\,Value \pm MOE \tag{14}$$

This calculation yields the upper and lower bounds of the confidence interval, which represent the range within which the true value is expected to lie with 95% confidence.

RESULTS AND DISCUSSION

Surface O_3 variations were studied during Vishu Festival-related fireworks on April 14 and 15 for 2019, 2020, 2021 and 2022 in Kannur city. The study period is divided into three sections viz. Days before Vishu (April 13), Vishu days (April 14, 15) and days after Vishu (April 16). Figure 1 shows the 24-hour variation in O_3 across the observation center for the days of the study period from 2019 to 2022. It can be seen that the diurnal variation in O_3 during these four years showed a similar pattern with different concentrations. No significant changes were noticed in O_3 in 2020 as fire extinguishing was banned during the nationwide lockdown to combat COVID-19 infection. The surface O_3 concentration is used to illustrate the basis of the data analysis, modelling and forecasting.

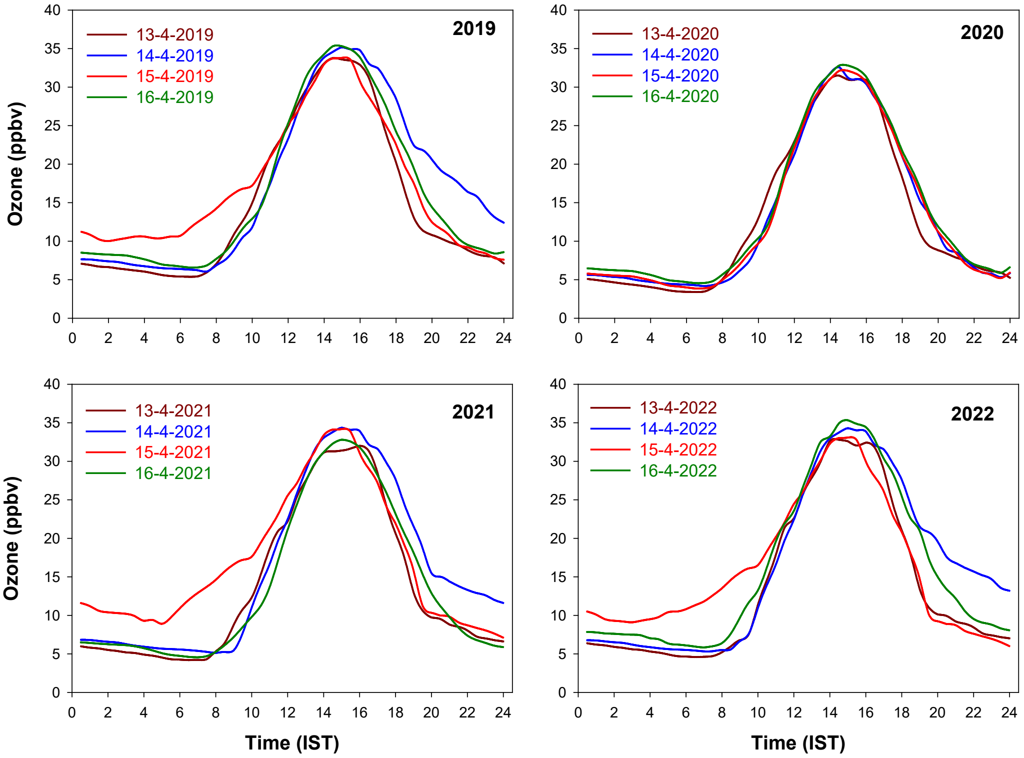

Fig. 1. Variation of surface O_3 during the period of observations

FORECASTING THE CONCENTRATION OF SURFACE O_3

Environmental data may be contaminated with errors and glitches that occur during the collection or transmission phase. They should therefore be refined before being processed by modelling or prediction systems. In this test, we examine the processes of removing outliers, fixing missing values, and smoothing. The O_3 and NO_2 data set over the days of the investigation period from 2019 to 2022 is pre-processed for the forecast and is shown in figure 2 (a & b).

The decomposition of surface O_3 data, such a s stochastic trends, seasonal changes, and random movements in the O_3 dataset for the study period is shown in the figure 3 (a, b,c,d) and for NO_2 is shown in the figure (e,f,g,h) respectively. O_3 time series chart is stochastic in nature. The inherent randomness of crackers, atmospheric conditions and wind patterns affect ozone concentrations. Additionally, uncertainties in data collection methods and the influence of random atmospheric variations contribute to the unpredictable behaviour observed in the O_3 time series. These factors may be influenced to form a combined result in the stochastic nature of the O_3 data. The stochastic trends, seasonal changes, and random movements in the O_3 data are made abundantly obvious from this figure (the second, third, and bottom panels of figure 3).

We use the slope of Sen for O_3 to determine the actual trend and find that the slope estimate is 0.0011. The estimated slope (0.11) is positive, so surface O_3 is trending up by 0.11 ppb. Strong seasonality is evident in both the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the surface O_3 and NO_2 data, consistent with the results presented in figure 4 (a and b). A seasonal trend is evident in both the ACF and PACF of the O_3 data, with most peaks occurring within the two confidence intervals. Thus, O^3 information can be considered static.

Formal tests such as Augmented Dickey-Fuller (ADF) as well as Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) are employed to ensure the stationary nature of O_3

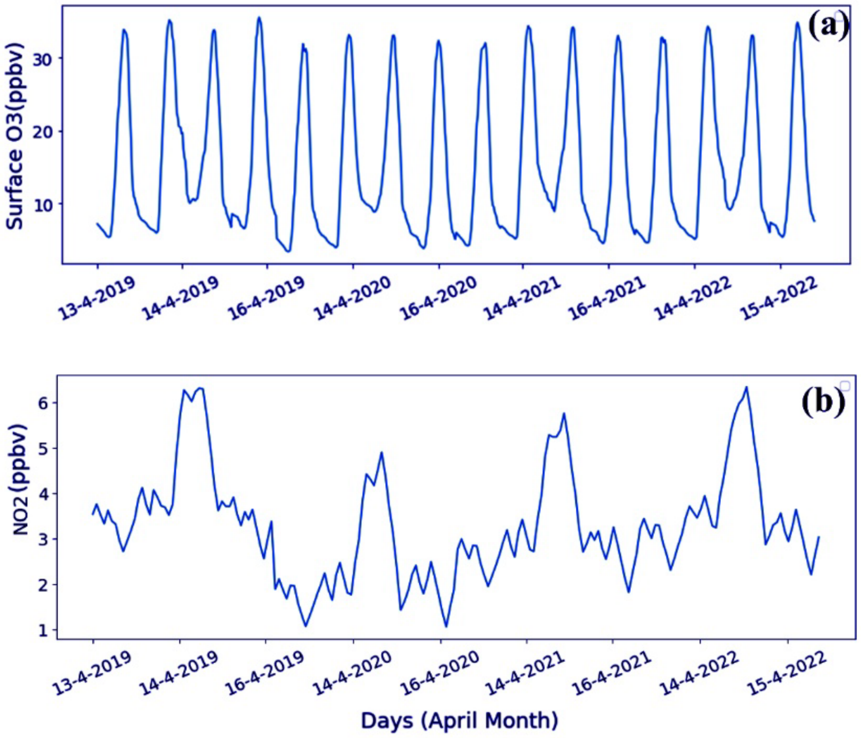

Fig. 2. Time series dataset of (a) surface O_3 (b) NO_2

data. Based on the estimated values of ADF, PP and their associated p-values, the O_3 time series data appear to be stable. Based on the estimated values of ADF and KPSS and their associated p-values, the NO_2 time series data appear to be stable. The statistical details are presented in Table 1.

Construct SARIMA (p,d,q)(P,D,Q) 12 models using the Box-Jenkins modelling approach. The O_3 data are stationary, so the values of D and d are '0'. As a result, the SARIMA (p, d, q) (P, D, Q) 12 model is transformed into the SARIMA (p, 0, q) (P, 0, Q)12 model. Significant peaks (Figure 5) were observed at lags 1, 2, 3, and 4 for ACF and lags 1 and 2 for PACF. The models under consideration are listed together with the results of the Akaike, Hannan-Quinn, and Bayesian information criterion tests are presented in In Table 2. The model SARIMA (1,0,2) (2,0,4,12) exhibits the minimum AIC, HQIC, and BIC values among the selected models. SARIMA (1, 0, 2) (2,0,4,12) is the best model that was ultimately chosen.

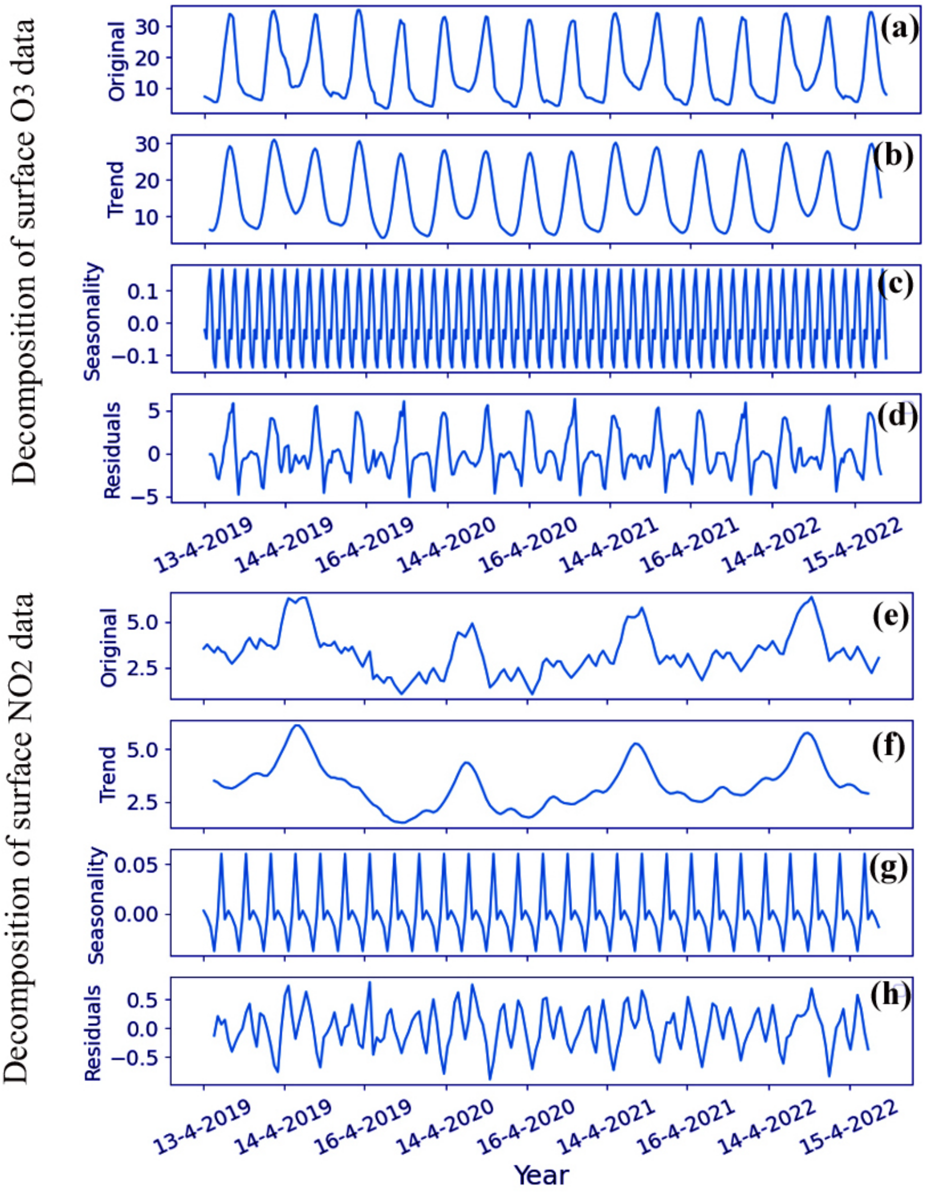

Fig. 3. Decomposition of (a,b,c,d) surface O_3 (e,f,g,h) NO_2 data for the study period

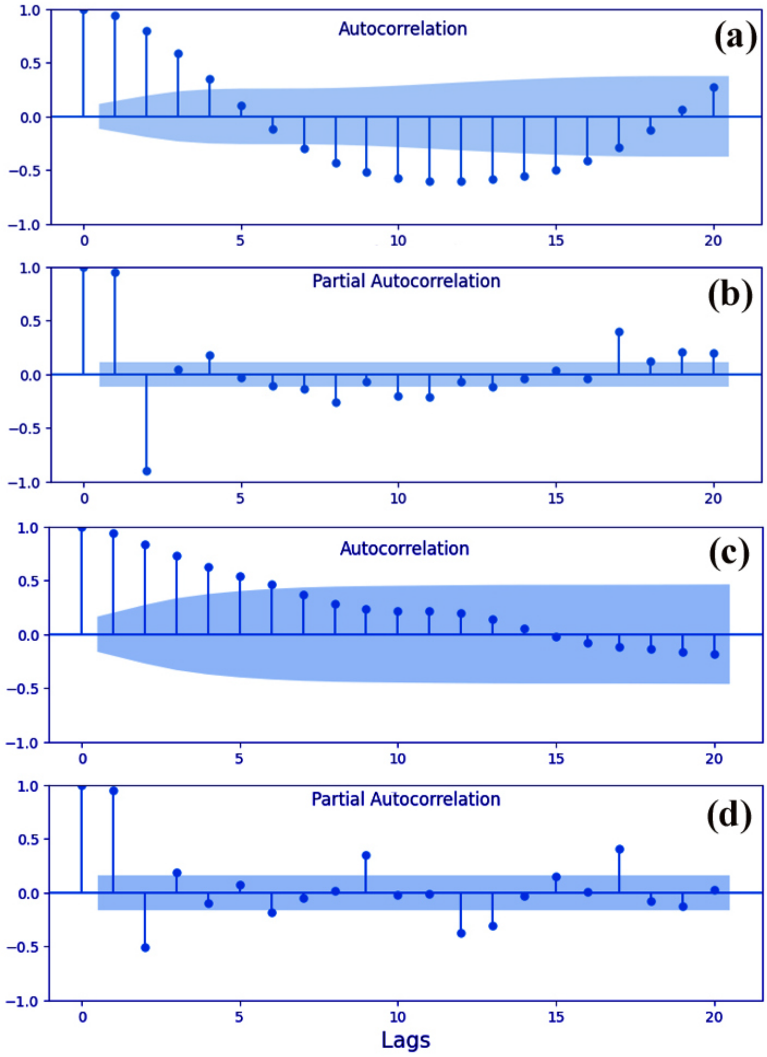

Fig. 4. (a) Autocorrelation function for surface O_3 (b) partial autocorrelations function for surface O_3 (c) autocorrelation function for NO_2 (d) partial function for NO_2

Table 1. Statistics table for both the ADF, PP and KPSS tests

Parameter	Name of test statistic	Calculate value	Lag order	*p*-value	Comment
O_3	ADF	-3.906	17	0.002	Stationary
	PP	-3.577	17	0.006	Stationary
NO_2	ADF	-3.709	3	0.004	Stationary
	PP	-3.021	20	0.126	Stationary
	KPSS	0.269	17	0.003	Stationary

The residuals are displayed in Figure 5(a) in which they seem to be random and not seasonal at all. Figure 5(b) shows that the kernel density estimate (KDE; orange curve) approximately coincides with N(0,1) (green curve). With a mean of zero and a standard deviation of one, the residuals follow a normal distribution. The red line and blue dots in Figure 5(c) represent residuals and normally distributed data with a mean of zero and a standard deviation of

Table 2. A summary of the criteria for selecting models for the various models

Model	AIC	BIC	HQIC
(1,0,2) (2,0,1,12)	906.827	931.833	916.876
(1,0,2) (2,0,2,12)	795.432	823.948	806.894
(1,0,2) (2,0,3,12)	753.645	785.302	766.387
(1,0,2) (2,0,4,12)	689.068	723.748	703.046

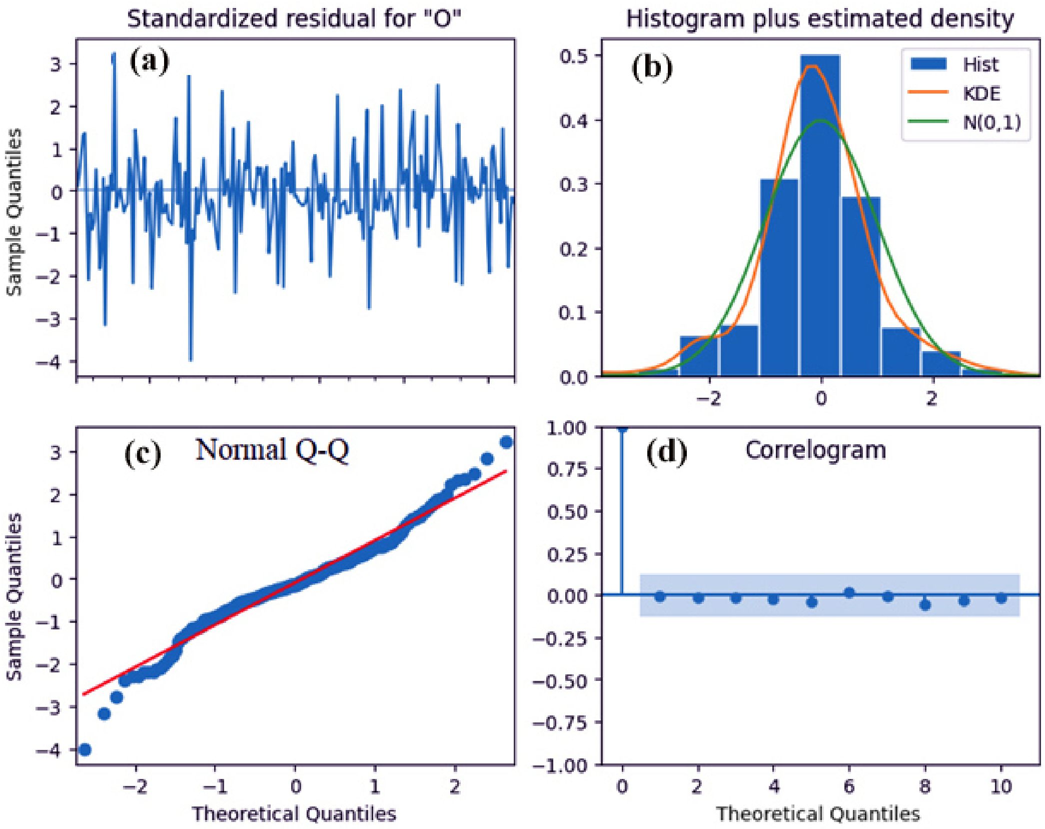

Fig. 5. Residuals plot for surface O_3. (a) Periodic residuals; (b) histogram of frequency distribution; (c) Q-Q plot; (d) autocorrelation

one, respectively. Q-Q plotting the residuals displays a linear trend. This results in a normal distribution for the residuals. Owing to its high level of accuracy in making predictions, this model offers a promising method for predicting the future. The autocorrelation in Fig. 5(d) illustrates the residuals from the original data are found to be distinct from the delayed data.

Forecasted values with 95% confidence limit using SARIMA (1,0,2) (2,0,4,12) model for surface O_3 and SARIMA (1,0,1) (4,0,4,12) for NO_2 is shown in figure 6 a and 6b.

The forecasted values shown in figure 7 are highlighted in red colour line. Shaded areas often represent a range of uncertainty, such as 95% confidence intervals around a prediction.

When considering LSTM, figure 7 shows line plots of training and evaluation loss values over several training periods. See loss curves that look similar to the above, but are not necessarily identical, because while performing the transformer model it needs to be started from scratch, and the training and evaluation loss values depend on how the model weights are set. But these loss curves show how the learning performance changes as the number of epochs goes up. They

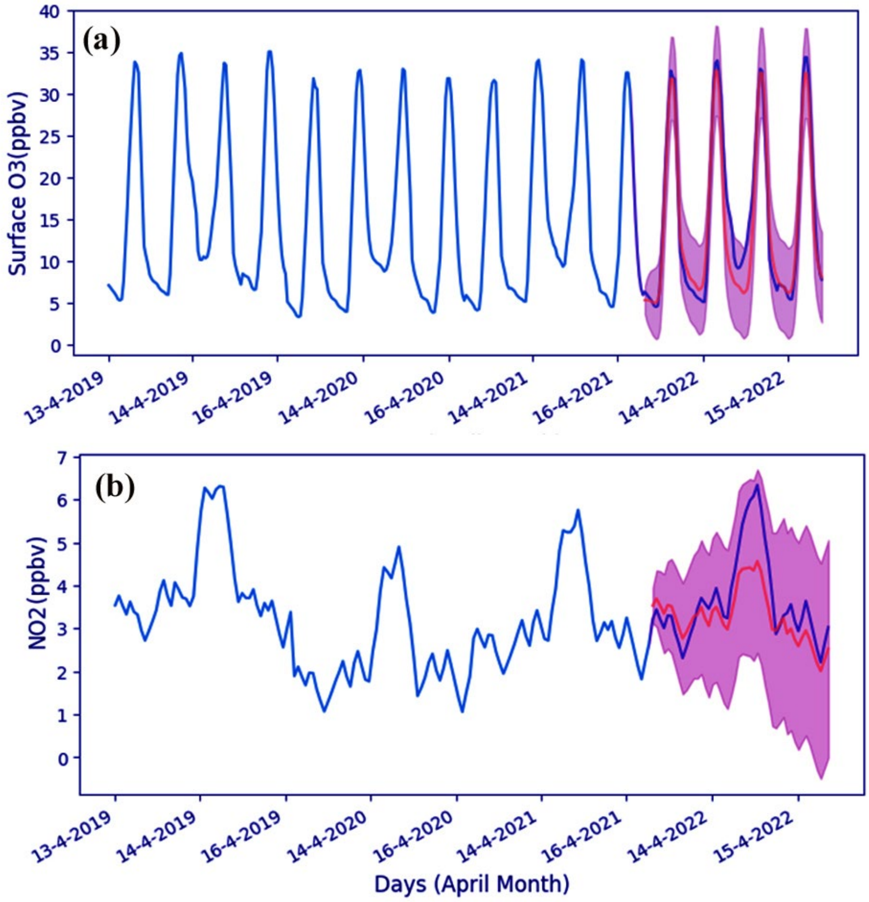

Fig. 6. SARIMA forecast of (a) surface O_3 (b) NO_2

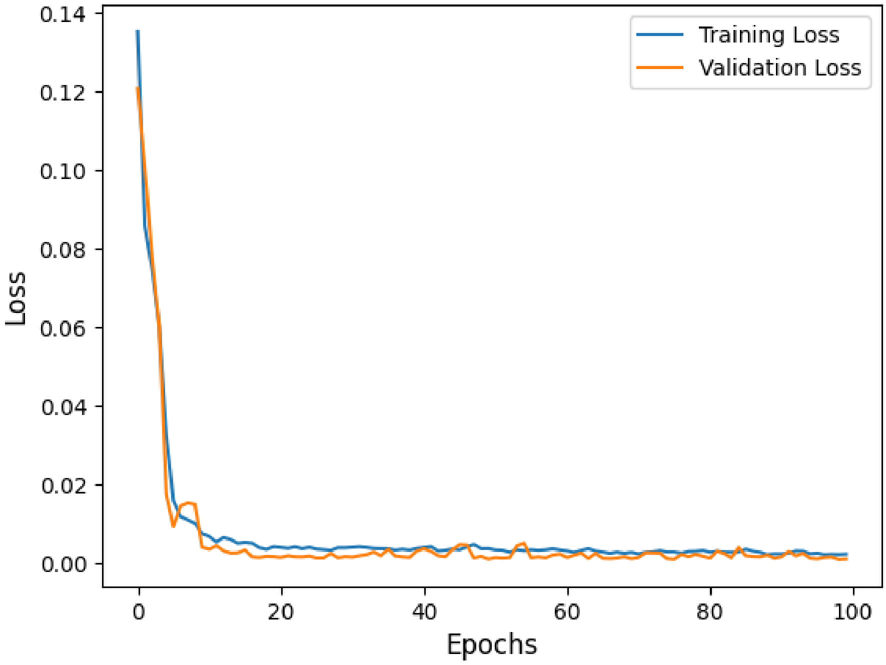

Fig. 7. Line plots of the training and validation loss values over several training epochs

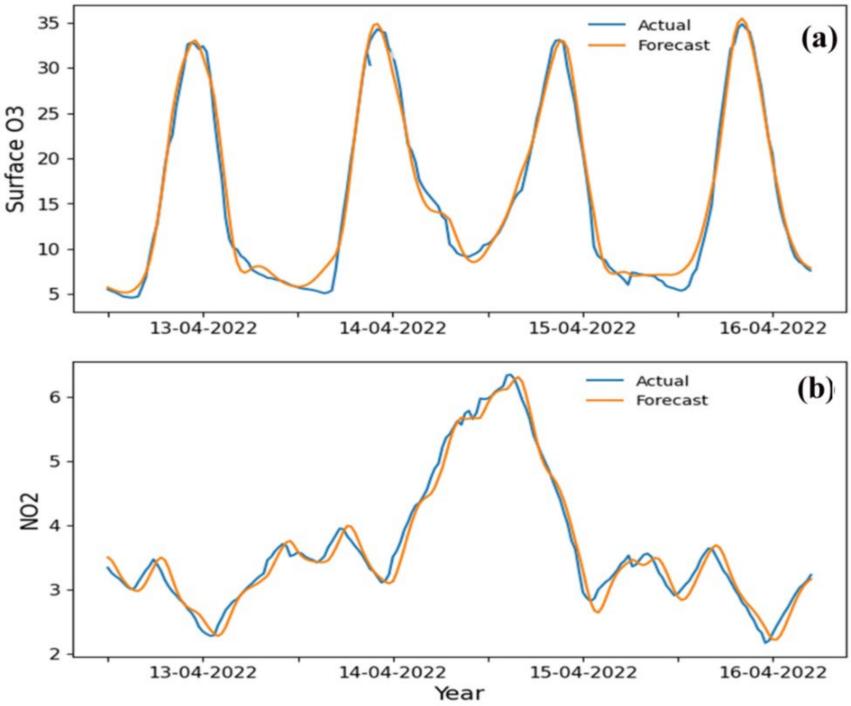

Fig. 8. LSTM for (a) surface O_3 (b) NO_2

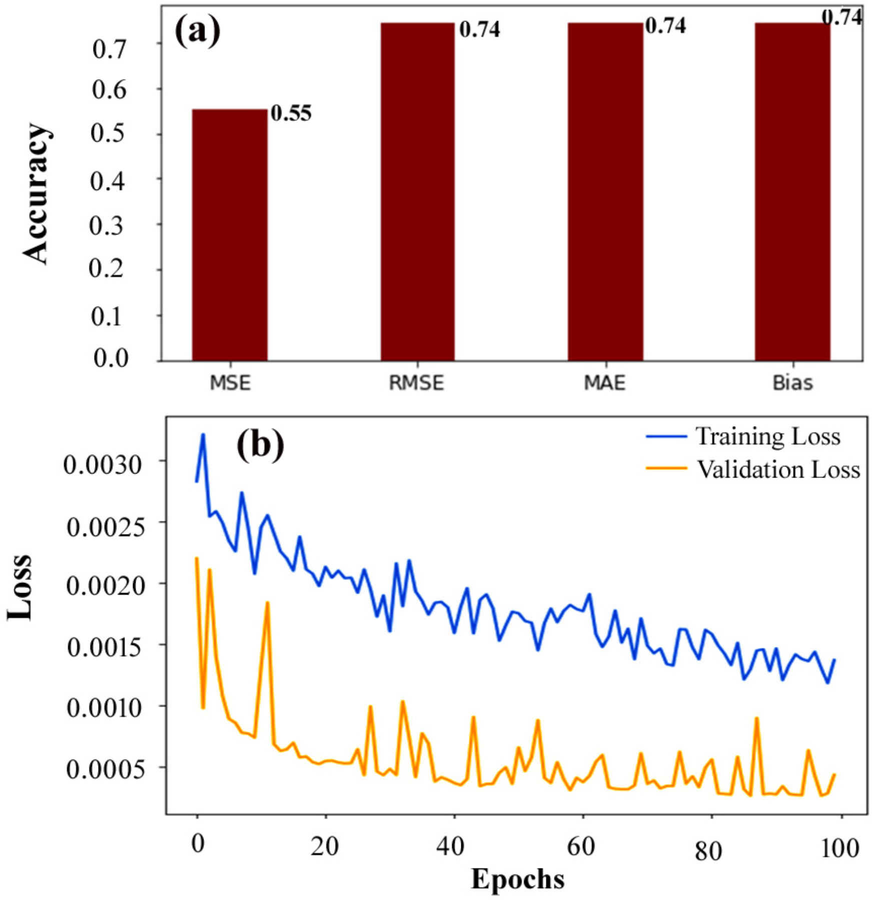

Fig. 9. (a) Performance measures (b) training and validation loss values over several training epochs

also tell us if there are any problems with the study that could cause the model to be under fit or over fit.

Figure 8(a) depicts the LSTM for surface O_3 and 8 (b) depicts the LSTM for NO_2 with an RMSE of 1.21. The sequence of values is critical when dealing with time series data. A simple approach for dividing an ordered dataset into train and test datasets. The split point index divides the data into training datasets that may be used to train our model with 75% of the observations and test the model with the remaining 25%. Finally, we may create predictions using the train and test dataset models to see how well the model performs. The predictions must be adjusted to match with the original dataset on the x-axis thanks to the way the dataset is produced. The data are shown after pre-processing, with the original dataset in blue, the predictions for the training dataset in green, and the training dataset in orange.

LSTM model shows high RMSE compared to stochastic model, due to the less number of input values while the LSTM model requires a large number of input values for better performance. Based on SARIMA model, figure 9 (a) shows the performance measures of the forecasting models. From the figure it's clear that the MSE is 0.055, RMSE is 0.74, Mean Absolute Error (MAE) is 0.74 and Bias (Forecast error) is 0.74. When considering LSTM, figure 7 shows line plots of training and evaluation loss values over several training periods. These loss curves show how the learning performance changes as the number of epochs goes up. They also tell us if there are any problems with the study that could cause the model to be under fit or over fit.

CONCLUSION

This is an attempt to forecast the O_3 concentration at Kannur from the previous data sets by using the Artificial Neural Networks model. An increase in ground-level O_3 and its precursor NO_2 has been detected during the fireworks associated with the Vishu festival, held every year on April 14 and 15 in Kerala. Thus, O_3 and NO_2 data were analyzed for four consecutive days from 13 to 16 April 2019 to 2022. Thereby, RNN and SARIMA methods are used successfully for the prediction of the surface O_3 and NO_2 concentration in the following years.

The selected model is highly suitable for forecasting time series. A comparison between the expected and observed values can be used accuracy of prediction. Underfitting or overfitting can also be avoided by performing this test. Predictive results from statistical tests are analyzed in detail. This model predicts surface O_3 based on historical data sets of O_3 at Kannur. The years 2019-2021 are used as the training dataset and 2022 as the test dataset. The predicted MSE and RMSE are 0.55 and 0.74, respectively. The MSE is fairly low and trending upward at 0.0011 ppb. Given the ambiguity of the growing trend, the chosen model has high prediction accuracy on the test set and may be utilized for future work. The result shows that the SARIMA predicted value (orange line) is within the confidence interval (Gray shading) and close to the actual value (blue line). The predicted MSE is 0.55, which is quite low. Compared to the SARIMA model's high RMSE value in LSTM, (due to the less number of input values), the LSTM model requires a greater number of input values are required for better performance. The anticipated outcomes are generally positive.

The SARIMA model is considered to be one of the best options for O_3 and NO_2 forecasting due to its ability to capture both the temporal and seasonal patterns in O_3 data. By incorporating autoregressive, differencing, and moving average components, SARIMA models can effectively model the time-dependent behavior of ozone concentrations, while the inclusion of seasonal parameters allows for the consideration of recurring patterns. This makes SARIMA a suitable choice for forecasting ozone levels, as it can account for both short-term fluctuations and long-term trends, providing valuable insights for environmental monitoring and decision-making processes. Based on the present study, the selected SARIMA model has sufficient predictive accuracy to predict future values.

The SARIMA models offer numerous merits over Recurrent Neural Networks (RNN) for the analysis of atmospheric O_3 and NO_2. SARIMA models explicitly capture the seasonal patterns in the data through the incorporation of seasonal components. They can effectively model and forecast time series with recurring patterns, such as weekly, monthly, or annual seasonality. RNN models, on the other hand, may struggle to capture and interpret complex seasonal patterns without additional preprocessing or modifications. SARIMA models provide interpretable results, allowing for a clear understanding of the underlying factors influencing O_3 and NO_2 levels. The model coefficients represent the autoregressive and moving average components, allowing users to understand the impact of different variables on the time series. This interpretability can be crucial for policy-making and identifying actionable strategies for air quality improvement. SARIMA models can work well with small datasets. They are less data-hungry compared to RNN models, which typically require a large amount of data to effectively learn the patterns and dependencies in the time series. SARIMA models can provide accurate forecasts even with limited historical data, making them suitable for situations where data availability is limited.

Furthermore, the appropriate models incorporating the Hindcast protocols is another work initiated to observe the variation between observed and predicted surface O_3 concentration in fireworks over the last decade and we expect a better scenario after analysing seasonal HYSPLIT trajectories, which consider the air mass movement.

ACKNOWLEDGEMENT

The authors are grateful to the Kerala State Council for Science Technology and Environment (KSCSTE) and Kerala State Pollution Control Board (KSPCB) for their assistance in researching air pollutants over Kerala region. Furthermore, the authors acknowledge their gratitude to three anonymous reviewers for their insightful feedbacks and critical comments to improve the manuscript in the present form.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Antony, E., N S Sreekanth., R K Sunil Kumar., & Nishanth T. (2021). *Data Preprocessing Techniques for Handling Time Series data for Environmental Science Studies,* International Journal of Engineering Trends and Technology, 69(5), 196-207.

Arputharaj, S., & Samuel Selvaraj, R. (2016). Prediction of surface ozone using arima, International Journal of Current Research, 41643-41646.

Asfaw, A., Simane, B., Hassen, A., & Bantider, A. (2017). Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: A case study in Woleka sub-basin. *Weather and climate extremes, 19*, 29-41.

Attri A.K., Kumar U., & Jain V.K. (2001). Formation of ozone by fireworks. Nature, 411, 1015.

Basurko EA., Anta A., Barron LJR., & Albizu M. (2007) In: Borrego C, Brebbia C (eds) Air pollution XV, WIT transactions on ecology and the environment, 101, 109–118.

Beldjillali., Hicham., Bachari,, Nour El Islam., & Lamri, Nacef. (2016). Prediction of ozone concentrations according the Box-Jenkins methodology for Assekrem area, Applied Ecology and Environmental Sciences, 4, 48-52.

C. Paoli., G. Notton., M. -L. Nivet., M. Padovani., & J. -L. Savelli. (2011). A Neural Network model forecasting for prediction of hourly ozone concentration in Corsica, 10th International Conference on Environment and Electrical Engineering, 1-4.

Capilla C. (2016). Int J Sustain Dev Plan, 11(4),558.

Chandra, S., Ziemke, J. R., & Stewart, R. W. (1999). An 11-year solar cycle in tropospheric ozone from TOMS measurements. Geophysical Research Letters, 26(2), 185-188.

Chowdhury, M. H., Mondal, S., & Islam, J. (2018). Modeling And Forecasting Humidity In Bangladesh: Box-Jenkins Approach. International Journal of Research -GRANTHAALAYAH, 6(4), 50–60.

Coman A., Ionescu A., & Candau Y. (2008). Hourly ozone prediction for a 24-h horizon using neural networks, Environ Model Software, 23(12), 1407-1421.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association, 74(366), 427–431.

Duenas C., Ferna´ndez M., Canete S., Carretero J., & Liger E. (2005). Stochastic model to forecast ground-level ozone concentration at urban and rural areas, Chemosphere, 61(10), 1379-1389.

Gideon Schwarz. (1978). Estimating the Dimension of a Model, Ann. Statist., 6(2), 461 – 464.

Hamzaçebi, Coşkun. (2008). Improving artificial neural networks' performance in seasonal time series forecasting. Inf. Sci., 178, 4550-4559.

Hiroo Hata., Kazuya Inoue., Hiroshi Yoshikado., Yutaka Genchi., & Kiyotaka Tsunemi. (2023). Impact of introducing net-zero carbon strategies on tropospheric ozone (O3) and fine particulate matter (PM2.5) concentrations in Japanese region in 2050, Science of The Total Environment, 891, 164442.

Kumar, K., Yadav, A. K., Singh, M. P., Hassan, H., & Jain, V. K. (2004). Forecasting Daily Maximum Surface Ozone Concentrations in Brunei Darussalam—An ARIMA Modeling Approach. Journal of the Air & Waste Management Association, 54(7), 809–814.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., & Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. Journal of Econometrics, 54, 159-178.

Li, C., Balluz, L. S., Vaidyanathan, A., Wen, X. J., Hao, Y., & Qualters, J. R. (2016). Long-Term Exposure to Ozone and Life Expectancy in the United States, 2002 to 2008. Medicine, 95(7), e2474.

Ljung, G.M., & Box, G.E. (1978). On a measure of lack of fit in time series models. Biometrika, 65, 297-303.

Mahiyuddin, W. R. W., Jamil, N. I., Seman, Z., Ahmad, N. I., Abdullah, N. A., Latif, M. T., & Sahani, M. (2018). Forecasting Ozone Concentrations Using Box-Jenkins ARIMA Modeling in Malaysia. American Journal of Environmental Sciences, 14(3), 118-128.

Maji, K. J., & Namdeo, A. (2021). Continuous increases of surface ozone and associated premature mortality growth in China during 2015-2019. Environmental pollution (Barking, Essex : 1987), 269, 116183.

Monks P.S., Archibald A.T., Colette A., Cooper O., Coyle M., Derwent R., Fowler D., Granier C., Law K.S., Mills G.E., Stevenson D.S., Tarasova O., Thouret V., & Schneidemesser E. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer, Atmos. Chem. Phys., 15, 8889–8973.

Nariaki Sugiura. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections, Communications in Statistics - Theory and Methods, 7(1), 13-26.

Nishanth, T., Ojha, N., Kumar, M.K.S., & Naja, M. (2012). Influence of solar eclipse of 15 January 2010 on surface O_3. Atmospheric Environment, 45, 1752-1758.

Pawlak I., Fernandes A., Jarosławski J., Klejnowski K., & Pietruczuk A. (2023). Comparison of 24 h Surface Ozone Forecast for Poland: CAMS Models vs. Simple Statistical Models with Limited Number of Input Parameters. Atmosphere. 14(4), 670.

Phillips, Peter., & Perron, Pierre. (1986). Testing for a Unit Root in Time Series Regression. Cowles Foundation, Yale University, Cowles Foundation Discussion Papers. 75.

Resmi CT., Nishanth T., Satheesh Kumar MK., Balachandramohan M,. & Valsaraj KT. (2020). Annular Solar Eclipse on 26 December 2019 and its Effect on trace pollutant concentrations and meteorological parameters in Kannur, India: a Coastal City, Asian J Atmos Environ, 14, 289–306.

Samuel Selvaraj, R., Sachithananthem C.P., & K.Thamizharasan. (2013). Modeling and Predicting Total Ozone Column and Rainfall in Kodaikanal, Tamilnadu By Arima Process, International Journal Of Engineering And Computer Science, 2(8), 2521-2526.

Sen, P.K. (1968) Estimates of the Regression Coefficient Based on Kendall's Tau. Journal of the American Statistical Association, 63, 1379-1389.

Sepp Hochreiter., & Jürgen Schmidhuber. (1997). Long Short-Term Memory. Neural Comput, 9(8), 1735–1780.

Xu, W., Lin, W., Xu, X., Tang, J., Huang, J., Wu, H., & Zhang, X. (2016). Long-term trends of surface ozone and its influencing factors at the Mt Waliguan GAW station, China – Part 1: Overall trends and characteristics, Atmos, Chem. Phys., 16, 6191–6205.

Xu, W., Xu, X., Lin, M., Lin, W., Tarasick, D., Tang, J., Ma, J., & Zheng, X. (2018). Long-term trends of surface ozone and its influencing factors at the Mt Waliguan GAW station, China – Part 2: The roles of anthropogenic emissions and climate variability, Atmos. Chem. Phys., 18, 773–798.

Zhang, A., Fu, M., Feng, X., Guo, J., Liu, C., Chen, J., Mo, J., Zhang, X., Wang, X., Wu, W., Hou, Y., Yang, H., & Lu, C. (2023). Deep Learning-Based Ensemble Forecasts and Predictability Assessments for Surface Ozone Pollution. Geophysical Research Letters, 50(8), e2022GL102611.