Air Quality, Atmosphere & Health (2019) 12:899–908
https://doi.org/10.1007/s11869-019-00696-7

# Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India

Mrigank Krishan[1] · Srinidhi Jha[1] · Jew Das[1] · Avantika Singh[2] · Manish Kumar Goyal[1] · Chandrra Sekar[3]

Received: 11 February 2019 / Accepted: 5 April 2019 / Published online: 17 April 2019


## Abstract

Nowadays, monitoring and prediction of air quality parameters are becoming significantly important research topics in the context of increasing urbanization and industrialization. Therefore, efficient modelling of air quality parameters is essential because such an approach would enable to identify the existing and forthcoming implication of air pollution. In recent years, sharp rise in air pollution levels in Indian National Capital Territory of Delhi (NCT-Delhi) has made it the most polluted city of the world. Machine learning approaches are considered as an efficient and cost-effective method to model the air quality parameters and are widely used. However, current methods fail to incorporate long-term dependencies arising due to complex interaction of natural and anthropogenic factors. The present study is mainly aimed at predicting $O_3$, $PM_{2.5}$, $NO_x$, and CO concentrations at a location in NCT-Delhi using the long short-term memory (LSTM) approach, which is considered as more efficient over other deep learning methods. Factors and parameters such as vehicular emissions, meteorological conditions, traffic data, and pollutant levels are employed in five different combinations. Performance evaluation of LSTM algorithms for hourly concentration prediction is carried out during 2008–2010, and it is found that LSTM models efficiently deal with the complexities and is immensely effective in ambient air quality forecasting. This paper can be considered as a significant motivation for carrying research on urban air pollution using latest LSTMs and helping the government and policymakers a better forecasting methodology for planning measures to curb ill impacts of degrading air quality.

**Keywords** Air pollution · Machine learning · Deep learning · LSTM · NCT-Delhi

## Introduction

According to air pollution program of the World Health Organization (WHO), 91% of the world's population inhales polluted air, and about 4.2 million deaths occur every year because of ambient air pollution. It reveals that air pollution is causing more than one-third deaths from lung cancer, strokes, and chronic respiratory implications. India, a developing country, is largely dependent on its cities for growth and faces huge challenges in maintaining healthy air quality in cities due to industrialization, vehicles, and other anthropogenic sources (Ghasemi and Amanollahi 2019). It is expected that 40% of India's population will be residing in cities by the end of 2030 as reported by Press Trust of India (PTI 2018). As shown by several air studies monitoring studies, India's cities are witnessing world's worst form of atmospheric pollution because of a large number of emission sources, pollutant transport, high emission rates, and unfavorable emission sources (Guttikunda et al. 2014). Air quality is becoming a matter of concern in Indian cities due rising particulate matter (PM), carbon monoxide (CO), nitrogen oxides ($NO_x$), ozone ($O_3$), and sulfur dioxide ($SO_2$). According to WHO Global Urban Ambient Air Pollution Database, ten Indian cities were featured in the list of cities with worst $PM_{2.5}$ pollution (WHO 2018). The air pollution levels in the Indian National Capital Territory of Delhi (NCT-Delhi) began to exceed most other cities in developing nations since the early 1990s. There are many distinctive factors which impact the air pollution levels in NCT-Delhi such as unrestrained emission sources in its surrounding area where city regulations are either not applicable or strictly abided, a large number of uncontrolled sources within the city,

✉ Jew Das
jewdas05@gmail.com

[1] Discipline of Civil Engineering, Indian Institute of Technology Indore, Indore, Madhya Pradesh, India

[2] School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Mandi, Himachal Pradesh, India

[3] Department of Civil Engineering, Dr. Ambedkar Institute of Technology, Bengaluru, Karnataka, India

unfavorable local meteorological conditions such as extreme summers and extreme winters which govern the particle suspension in air, periodic agricultural pollutant transport from outskirts (Guttikunda and Gurjar 2012), feedback of ozone as local background, and the effect of travelling precursor pollutants from subcontinent as well as intercontinental sources (West et al. 2009).

Due to the complexity of natural and anthropogenic factors, prediction of air pollution has become an increasingly difficult task. Several approaches have been developed to model the air quality parameters considering transport emissions, meteorological conditions, and traffic flow. Generally, air quality models are based on either the deterministic analysis of fundamental atmospheric processes or the emissions or statistical models, which comprise the analysis of semi-empirical statistical associations among variables. Our inability to measure the atmospheric and boundary conditions and the intrinsic unpredictability of atmospheric motion prohibits us from considering any metrological phenomenon as fully deterministic. Such models may take into account a large number of parameters, however, still being influenced by many uncertainties in the model inputs, model parameterizations, and numerical formulations (Mallet and Sportisse 2008). On the other hand, the stochastic approaches involve many assumptions related to distribution and randomness. For instance, Milionis and Davies (1994) stated that statistical models might be able to incorporate the nonlinearity associated with the data that, however, may require basic assumption that it follows normal probability distribution and so on.

Existing researches in statistical modelling suggests that previously discussed methods have limited success in providing an accurate prediction of air quality parameters. On the other hand, machine learning techniques which have evolved over half a century have accomplished huge success in multiple areas. These methods do not necessarily require prior assumptions relating to data distribution and can perform self-formulation of the most efficient model by minimizing the cost function. With the advent of new technology and availability of larger data sets, machine learning techniques are gaining popularity in air quality prediction. Kalapanidas and Avouris (2001) deduced the impact of meteorological parameters such as precipitation, wind, temperature, humidity, and solar radiation and characterized air pollution into different classes (low, medium, high, and alarming) by employing a lazy learning method. Athanasiadis et al. (2003) used the $\sigma$-fuzzy lattice neurocomputing classifier to simulate and classify ozone levels (high, medium, and low) by analyzing the metrological data and other pollutants such as $NO_2$, NO, and $SO_2$. Jain and Khare (2010) performed the prediction of carbon monoxide concentration using adaptive neuro-fuzzy model with prediction accuracy of 89 to 93%. Kurt and Oktay (2010) incorporated the geographic relationship into artificial neural network (ANN) model and forecasted daily concentrations of CO, $SO_2$, and $PM_{10}$. It is important to note that the transition from regression to classification is not always efficient as it poorly deals with the magnitude of extreme values in the data. Fu et al. (2015) introduced a method of the ray model and rolling mechanism and to improve the simulation of traditional FFNN models. Ni et al. (2017) analyzed $PM_{2.5}$ concentration data around Beijing to compare a variety of statistical models, and their results reveals that linear regression models, in some cases, produce better outputs. Recently, recurrent neural networks (RNNs) which are considered more advantageous over ANNs because of its ability to model the interrelated sequence of time series also have been utilized in modelling air quality. For example, Kim et al. (2010) modeled the indoor air quality parameters using neural networks (NN), multiple linear regression (MLR), and recurrent neural network (RNN) and demonstrated that RNN model performance is better and has the capability to model such dynamic nonlinear systems. Due to the dynamic nature of RNNs, they can provide correct one-step and furthermore multi-step forward predictions. Fan et al. (2017) developed a spatiotemporal forecast framework for air quality modelling using deep RNNs and missing data processing algorithms. They further trained gradient boosting decision trees (GBDT) and deep feed forward neural networks (DFNN) as baseline models and observed that both DFNN and GBDT were outperformed by proposed deep RNN model. Theoretically, RNNs are capable of handling long dependencies in time series; however, sometimes, it may be possible that the gap between the point of information extraction and relevant information becomes very large. Unfortunately, in practicality, RNNs are rarely able to learn from such situations. Bengio et al. (1994) discussed the detailed explanations of these limitations. Briefly, the long-term dependency problem can further be elaborated as, when a larger network through time is encountered, the gradient decays rapidly during back propagation, hence, training RNNs comprising of long unfolding becomes difficult. However, long short-term memory (LSTM) networks are special types of RNNs, which are efficient in learning long-term dependencies. LSTMs evade the gradient decay issue by forming cell states through time, which allow the gradient to flow freely backward in time. Formulated by Hochreiter and Schmidhuber (1997), LSTMs were further modified and improved for more effective performance (Mikolov et al. 2014). Lately, LSTMs are successfully applied in variety of fields such a robot control (Mayer et al. 2008), speech and language learning (Graves et al. 2013), finance (Bao et al. 2017), biological sciences (Sønderby et al. 2015), hydrology, and water resources (Zhang et al. 2018). In the area of air quality monitoring, Pardo and Malpica (2017) employed LSTMs to predict

hourly forecasts of $NO_2$ levels up to 24 h in Madrid, Spain. In another study, it was observed that LSTMs were able to learn from spatial correlations and performed better modelling of regional air pollution (Li et al. 2017). Indoor air quality prediction using LSTM's capability of analyzing time series sensor data was attempted which provided good results (Ahn et al. 2017). Recently, Zhong et al. (2011) used RNNs with LSTM units combining input of air pollution and meteorological data in Daegu, Seoul, Beijing, and Shenyang cities and pointed out the importance of incorporating the meteorological information into modelling processes. The prediction accuracy of different configurations was compared, and they further suggested integrating multiple layers when predicting far time steps ahead. These results encouraged us to explore the possibilities of LSTM-based prediction of the air quality parameters in Indian capital city of Delhi.

The contributions from this paper are as follows: (i) to the best of our knowledge, ours is the first attempt to forecast air pollution parameters in any Indian city using LSTM and (ii) the proposed model outperforms the traditional methods. The following sections detail the obtained data for the analysis, development of the proposed model, and finally the discussion of the outcomes from the investigation.

## Materials and methods

### Study area

The National Capital Territory (NCT) of Delhi extends to around 1500 $km^2$ of area, including areas from the states of Haryana, Rajasthan, and Uttar Pradesh. The current population of Delhi is 18.6 million, and according to United Nation's World Urbanization Prospects (2018), it is expected to grow close to the population of world's largest city Tokyo by the end of 2030. Although, a number of emission control measures have been implemented to curb the air pollution; however, none of them have been much productive as the concentrations of major pollutants particulate matter, sulfur dioxide, and nitrogen oxides continue to rise (Gurjar et al. 2016). In this study, we selected location in northern part of Delhi based on difference in street configuration and traffic density. Our location is situated near Income Tax Office (ITO), where an air pollution monitoring station is maintained by central pollution control board (CPCB). Figure 1 shows our study area and the location of air pollution monitoring station state of traffic recorded on Aug. 11, 2018. The region comprises of commercial and administrative buildings, which attract a large volume of traffic. Several reports mention the large flow of

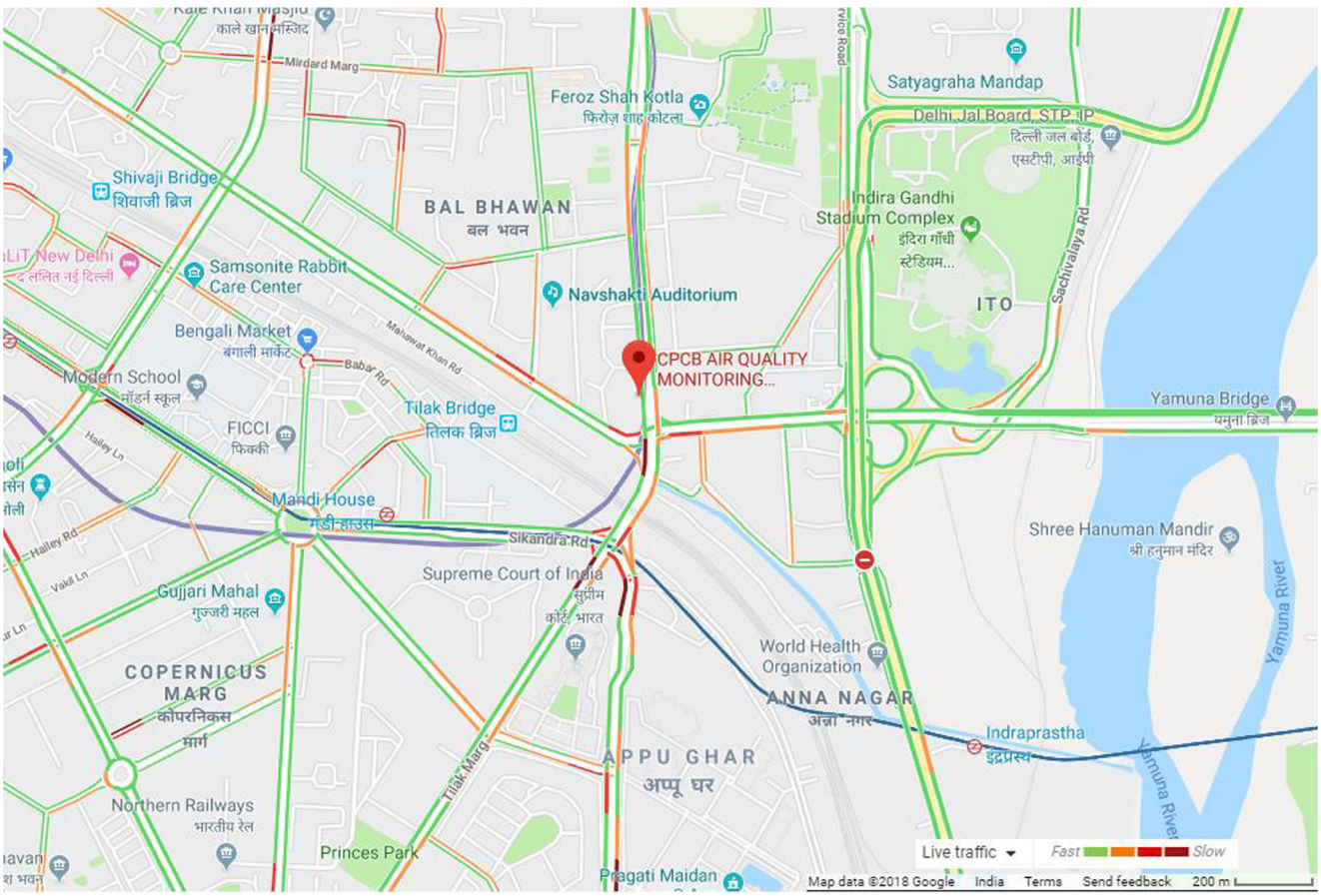


**Fig. 1** Study area and the location of air pollution monitoring station

vehicle across this intersection every day. An estimate by City Development Plan (CDP) shows that about 110,000–166,175 pass through this region each day, approximately 8–12% of them during peak timings (9:00–11:00 a.m. and 5:00–7:00 p.m.) (Chugh et al. 2012). Approximately 91% are light vehicles, two wheelers, and three wheelers, and approximately 9% are heavy-duty vehicles (Gokhale and Pandian 2007).

## Data sets

### Meteorological data

The meteorological data used in this study was obtained from Safdarjung meteorology monitoring station located at around 6 km from the study area. Table 1 shows the list of hourly Indian Meteorological data for the period 2008–2010 used in the study with their descriptive statistical measures. The mixing height data for analysis was used from the study carried out by Guttikunda and Calori (2013). We also estimated the hourly stability categories by employing Pas quill-Gifford (PG) stability scheme (Schnelle and Dey 2000).

### Transport emissions

Gurjar et al. (2004) discussed the method for total vehicular emission, which suggests that the emissions are based on emission factor for the type of vehicle, the distance travelled by a specific vehicle, the total number of vehicles, and their distribution in the type of fuel used. This can be mathematically represented as follows:

$$E_i = (Veh_m \times D_n) \times E_{i;j\ \mathrm{km}} \quad (1)$$

where $E_i$ = emission of compound$i$; $Veh_m$ = total number of vehicles per type $n$; $D_j$ = distance traversed per different vehicle type $j$; and $E_{i;\,j\ \mathrm{km}}$ = emission of compound $i$, vehicle type $j$ per kilometer travelled.

### Traffic data

We obtained the data of different type of vehicles registered during the 2000 to 2010 from Transport Department, New Delhi. The hourly average vehicle data for the period of 10 years 2000 to 2010 crossing the ITO intersection was acquired from Central Road Research Institute (CRRI), New Delhi. Using the Automative Research Association of India-ARAI (2007) guidelines, we estimated the emission factors for the vehicles. The classification consists of light-duty gasoline driven, light-duty diesel driven, CNG-driven vehicles, two wheelers, and LPG-driven vehicles.

### Air quality parameters

Hourly average ambient CO, $PM_{2.5}$, $NO_x$ and HC concentration data at the ITO intersection for the period of 3 years (i.e.,

**Table 1** List of variables and their descriptive statistics used in this study

| Input variable | Minimum | Maximum | Average | Standard deviation |
|---|---|---|---|---|
| Atmospheric pressure (millibars) | 914.04 | 999.39 | 982.87 | 6.90 |
| Relative humidity (%) | 8.80 | 431.30 | 64.13 | 22.33 |
| Temperature (°C) | 4.06 | 44.40 | 25.30 | 8.30 |
| Wind speed (m/s) | 0.00 | 36.53 | 1.46 | 1.72 |
| Wind direction (°) | 22.50 | 360.00 | 239.74 | 103.86 |
| Cloud cover (number) | 0.00 | 9.00 | 2.52 | 2.65 |
| Visibility (number) | 14.00 | 99.00 | 94.65 | 1.12 |
| Sunshine (h) | 0.00 | 11.60 | 5.80 | 3.21 |
| Rainfall (mm) | 0.00 | 859.30 | 2.81 | 24.30 |
| Stability class (number) | 1.00 | 6.00 | 4.26 | 1.58 |
| Solar insolation (Langley/h) | 0.00 | 114.72 | 15.88 | 23.09 |
| Mixing Height | 14.30 | 476,144 | 881.77 | 783.96 |
| Diesel-powered vehicles (number/h) | 21.00 | 1541.00 | 640.56 | 386.79 |
| Gasoline-powered vehicles (number/h) | 316 | 14,364 | 4925.3 | 3370.77 |
| CNG-powered vehicles (number/h) | 183.00 | 5487.00 | 2109.44 | 1369.48 |
| LPG-powered vehicles (number/h) | 1.00 | 79.00 | 32.81 | 19.83 |
| HC emissions (μg/s − m) | 158.56 | 6508.55 | 2226.22 | 1588.02 |
| CO emissions (μg/s − m) | 375.25 | 12,607.37 | 4842.82 | 3211.60 |
| PM emissions (μg/s − m) | 3.49 | 132.64 | 47.43 | 31.47 |
| $NO_x$ emissions (μg/s − m) | 65.59 | 2144.19 | 931.65 | 571.01 |

2008–2010) were obtained from CPCB, New Delhi. The height of the receptor at the station was 3.5 m above ground level and 3 m from the curbside.

The details of the data used in the present study and their descriptive statistics are presented in Table 1.

Appropriate training and testing data choice are crucial in evaluating the performance of the correct model. There is no thumb rule for partitioning the full data into training and testing sets. Each of our data set listed in the Table 1 consist of 26,305 hourly measurements. We have used 80% of the data set as training subset and 20% as the test subset to allow sufficient training of models. It is evident from Table 1 that our data consists of various types of variables with varying mean and standard deviation. The high values in data sets (e.g., number of vehicles) can influence the model and slow down the learning process. Hence, we normalized the data set between 0 and 1 to avoid numerical complexities. Moreover, the values close to 0 is replaced by smallest possible positive real number. Further, to predict the hourly concentrations of air quality parameters, special care was taken while combining the input variables. Table 2 shows the details of model used in this study with their input variable combinations. It is worth mentioning that the selection of input data is carried out using the following conditions: (i) meteorological variables play a significant role in transporting or spreading the roadway and industrial pollutants (Srivastava and Jain 2005); (ii) traffic related data are equally important to map the air quality parameters (Briggs et al. 2000); and (iii) vehicular emission data are equally important in modelling and cannot be ignored completely (Sekar et al. 2016b) (as discussed in "Transport emissions"). The models we employed are based on the suitable combination of the three above data types.

**Table 2** Details of input variables for different models

| Model number | Input variable combinations |
|---|---|
| Model 1 | Pressure, temperature, relative humidity, wind direction, wind velocity, visibility, cloud cover, rainfall, sun shine, solar insulation, stability class, mixing height, gasoline powered vehicles, diesel powered vehicles, CNG powered vehicles, LPG powered vehicles, CO, $NO_x$, HC, PM, (all) |
| Model 2 | Atmospheric pressure, temperature, relative humidity, wind speed, wind direction, visibility, cloud cover, rainfall, sunshine, solar radiation, stability class, mixing height (12) |
| Model 3 | Atmospheric pressure, temperature, relative humidity, wind direction, wind speed, visibility, cloud cover, rainfall, sunshine, solar radiation, stability class, mixing height, gasoline, diesel, CNG, LPG 16 |
| Model 4 | Atmospheric pressure, temperature, relative humidity, wind direction, wind speed, visibility, cloud cover, rainfall, sunshine, solar radiation, stability class, mixing height, CO, $NO_x$, HC, PM. |
| Model 5 | Gasoline, diesel, CNG, LPG, CO, $NO_x$, HC, PM, |

## LSTM model introduction

Figure 2 explains the working of LSTM model we have employed in this study. In comparison to RNNs, a standard LSTM neural network model has exclusive "cell states" which stores memory throughout the process. Every block consists of one or more self-recurrent cells and three other (input gate, forget gate, and output gate) units that support the continuous process of read, write, and reset operations for the cells.

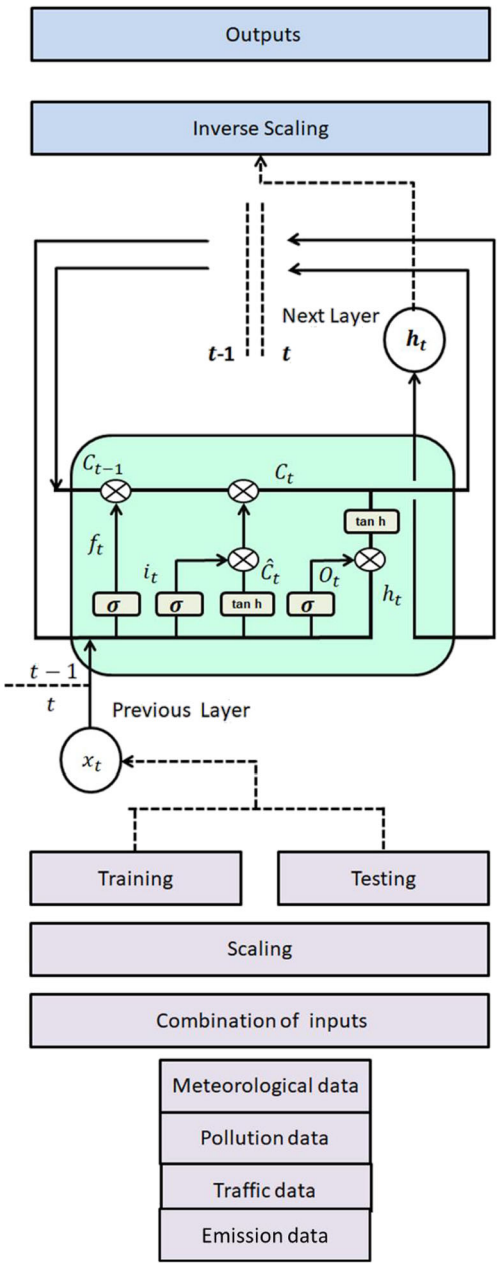


**Fig. 2** Flow diagram of methodology employed in this study with a standard LSTM network

The first step involves the identification of information that is not required in next steps and needs to be forgotten before passing to cell state. This is achieved by passing the information arriving from previous layer and the input through a sigmoidal layer which converts the resulting vectors $f_t$ values in the range of (0, 1). The output being released from the forget gate can be represented as can be represented as follows:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{2}$$

where $\sigma$ is the logistic sigmoid function and $W_f$ and $U_f$ are the adjustable weights or learning rates for this gate accompanied with $b_f$, the bias vector. $h_{t-1}$ is the hidden state which is initialized by a set of zero vectors with a user defined length. In the next step, the cell state information is updated by $\hat{c}$ which is a combination of input from the last hidden state $h_{t-1}$ and present input $x_t$, which can be given by the following equation:

$$\hat{c} = \tanh\left(W_{\hat{c}x_t} + U_{\hat{c}} h_{t-1} + b_{\hat{c}}\right) \tag{3}$$

where $\hat{c}$ is a set of values in the range of (− 1, 1), tanh (.) is hyperbolic tangent, and $U_{\hat{c}}$, $W_{\hat{c}}$ and $b_{\hat{c}}$ are weight and bias parameters respectively. Along with this vector, another set of values passing through sigmoidal function is calculated, which considers the degree to which input information is to be engaged in the cell state. This is known as the input gate and for a given time step, it can be represented as

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{4}$$

where, $W_i$, $U_i$, and $b_i$ are the weight rates and bias vector for the input gate. The product of values at this step and the previous step is combined to update the cell state using the following equation:

$$c_t = f_t \times c_{t-1} + i_t \times \hat{c} \tag{5}$$

Since both the vectors $f_t$ and $i_t$ are in the range (0, 1), the previous equation intends to forget or retain information supplied by $c_{t-1}$. Similarly, the vector $i_t$ decides which new information contained in $\hat{c}$ will be merged in cell state (values of $i_t$ tending to 1) and which will not be merged (values of $i_t$ tending to 1). Similarly, $i_t$ decides which values stored in $\hat{c}$ will be merged to the cell state and which will be left out. The final gate, which is the output gate, manages the information of the cell state $c_t$ that is provided to the new hidden state $h_t$. The output gate is computed as

$$o_t = \sigma\left(W_o x_t + U_o h_{(t-1)} + b_o\right) \tag{6}$$

where $o_t$ is a set of values in the range (0, 1) and $U_o$, $W_o$, and $b_o$ are a set of learning parameters, defined for the gate. From this vector, the new hidden state $h_t$ is estimated by combining the outputs of Eqs. 5 and 6 as follows:

$$h_t = \tanh(c_t) \times o_t \tag{7}$$

The cell state ($c_t$) governs the effective learning of the long-tern dependencies of network. Because of the simplistic linear interaction with the next LSTM cells, it can retain the information as it is over a very long period. When the network is trained, this property of LSTM helps to avoid the problem of exploding or vanishing gradient. Further, the length of cell state and the hidden state vectors in the LSTM can be chosen independently. There are several other parameters, which influence the working and mechanism of the LSTM neural networks such as the length of input sequence or the batch size, learning, rate and epochs. In this study, to obtain the best performance parameters, we carried out computations with different sets of these parameters for a double-layered LSTM network. Initially, we varied the learning rate, epochs, and batch size as 0.01–1, 1–100, and 50–1000 with respective increments of 0.1, 25, and 50. After the investigation of Nash–Sutcliffe efficiency (NSE) values, we narrowed our search to finer scale for epochs and batch size keeping learning rate constant 0.01 for all cases. Moreover, the hyper-parameter of the LSTM model was obtained through the grid search method and the best hyper-parameter was selected according to the overall accuracy of the model through the statistical efficiency tests ("Model evaluation parameters"). In other words, the hyper-parameter giving better simulated results was chosen for the study. During the modelling process, the overfitting of the model was also taken care by judging the performance of the model during both the training and testing sets. While simulation, it was observed that the loss function for training and testing decreases and stabilize around some point, which suggests that the model was good fit. Moreover, we considered a large number of simulations by varying the epochs, input, batch size, learning rate, and the ability of the model to capture the simulation outputs that were evaluated as compared to the observations during training and testing to minimize the chances of overfitting. In the present investigation, the time and space complexity was not computed; however, the complexity of LSTM learning with stochastic gradient descent optimization is O (1). Therefore, considering the total number of parameters (N), the complexity per time step is O (N). The value of $N$ can be computed as $N = n_c \times n_c \times 4 + n_i \times n_c \times 4 + n_c \times n_o + n_c \times 3$, where $n_i$ is the input nos., $n_c$ is the number of memory cells, and $n_o$ is the number of output units. The range of tested parameters in second step for predicting $O_3$, $PM_{2.5}$, $NO_x$ and CO is shown in Table 3.

**Table 3** The range of tested model parameters at finer scale

| | Initial value | | | | Final value | | | | Increment | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $O_3$ | $PM_{2.5}$ | $NO_x$ | CO | $O_3$ | $PM_{2.5}$ | $NO_x$ | CO | $O_3$ | $PM_{2.5}$ | $NO_x$ | CO |
| Epochs | 51 | 51 | 51 | 26 | 76 | 76 | 76 | 76 | 1 | 1 | 1 | 1 |
| Batch size | 50 | 50 | 50 | 50 | 650 | 250 | 350 | 200 | 25 | 15 | 25 | 15 |

## Model evaluation parameters

The efficiency of models is evaluated by performance indices such as NSE index, correlation coefficient (CC), percentage bias (PBIAS), and root mean square error (RMSE). The mathematical representation of the statistical parameters (NSE, CC, PBIAS, and RMSE) used in the study can be represented as follows:

1. Nash–Sutcliffe efficiency index

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(X_p - X_o\right)^2}{\sum_{i=1}^{N}\left(X_o - \overline{X}\right)^2} \tag{8}$$

2. Correlation coefficient

$$CC = \frac{\sum_{i=1}^{N}\left(X_o - \overline{X_o}\right)\left(X_o - \overline{X_p}\right)}{\left[\sum_{i=1}^{N}\left(X_o - \overline{X_o}\right)^2 . \sum_{i=1}^{N}\left(X_p - \overline{X_p}\right)^2\right]^{\frac{1}{2}}} \tag{9}$$

3. Percentage bias

$$PBIAS = \frac{\sum_{i=1}^{N}\left(X_o - \overline{X_p}\right)}{\sum_{i=1}^{N}\left(X_o\right)} \tag{10}$$

4. Root mean square error

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(X_p - X_o\right)^2}{N}} \tag{11}$$

where $X_o$=observed data; $X_p$ = predicted data; $N$ = total number of observations in data; predicted data, $\overline{X_o}$ = mean of observed data; and $\overline{X_p}$ = mean of the predicted data.

## Results and discussions

We tested five different combinations (models) of input parameters to simulate air quality. Initially, all test runs were performed on a coarser scale to get a broad idea about the behavior of model parameters. The preliminary investigation of model evaluation parameters suggested that models 4 and 5 are not efficient in forecasting our data sets. Hence, models 1, 2, and 3 were chosen and tested on a finer scale within the model parameters range as shown in Table 3. The epochs for $O_3$, $PM_{2.5}$, and $NO_x$ were varied from 51 to 76 with an increment of 1. However, for CO, the epoch range was 26 to 76 with an increment of 1. The initial value of batch size, which is the total number of data point, was kept 50 for all four cases and tested with an increment of 25, 15, 25, and 15 up to the final range of 650, 250, 350, and 200 respectively.

It should be noted that all training in the second step were made with the previously obtained best-fit learning rate of 0.01. We found that the best epoch size was 73, 72, 72, and 69 for $O_3$, $PM_{2.5}$, $NO_x$, and CO respectively. Similarly, model performance was optimum at a batch count of 75, 80, 175, and 95 respectively. We observed that the performance was hugely influenced by the type and number of input parameters we considered. The initial round of training suggests that the air quality parameters are essentially governed by the meteorological variables. Model 5, which comprised of eight input variables of different vehicle types and hourly ambient concentrations of CO, $NO_x$, HC, and PM, was thoroughly discarded. First model, which contained a full set of input variables combining meteorological data, traffic data, and emissions data, was found to be most suitable. Moreover, model 4 was also performed poorly possibly because of exclusion of traffic data. However, model 3 which also lack traffic and emission data were judged best for $NO_x$. Almaraz et al. (2018) reported that agriculture is the major source of $NO_x$. NCT-Delhi is badly affected by agricultural pollution from the nearby states. Moreover, the travelling of these agricultural pollutants is affected by the weather conditions and, hence, it may be the reason for obtaining optimum performance for $NO_x$ with the meteorological variables. It is worth mentioning that in this study, we have not considered the pollutants due to the agricultural pollution. Model 1 was found to be best for predicting $O_3$ and $PM_{2.5}$ concentrations, whereas CO was best predicted by model 2.

### Training results

It is observed from Table 4 that LSTM models performed quite well for all the four variables. The range of NSE during training was 0.85–0.96. The best predicted variable was $PM_{2.5}$

**Table 4** The estimates of model evaluation parameters for training and testing sets

| Parameter | | NSE | CC | PBIAS | RMSE |
|---|---|---|---|---|---|
| $O_3$ | Training | 0.85 | 0.92 | 0.02 | 2.91 (μg/m$^3$) |
| | Testing | 0.86 | 0.93 | 0.03 | 2.71 (μg/m$^3$) |
| $PM_{2.5}$ | Training | 0.96 | 0.98 | −0.01 | 1.64 (μg/m$^3$) |
| | Testing | 0.94 | 0.97 | −0.06 | 1.84 (μg/m$^3$) |
| $NO_x$ | Training | 0.88 | 0.94 | 0.01 | 2.61 (ppb) |
| | Testing | 0.86 | 0.93 | 0.01 | 2.72 (ppb) |
| CO | Training | 0.90 | 0.95 | 0.00 | 1.50 (μg/m$^3$) |
| | Testing | 0.90 | 0.95 | −0.01 | 1.42 (μg/m$^3$) |

according to highest NSE value of 0.96. Least value of NSE was obtained as 0.85 in the case of $O_3$. The NSE, which determines the relative measure of residual variance (noise) noise as compared to measured data information (data variance), reveals that LSTM models have better efficiency in predicting the air quality parameters. Although NSE is widely accepted as an effective model evaluation parameter, still, it suffers from some limitations. The biggest shortcoming of the NSE is that the differences between the predicted values and observed values are estimated as squared values. As a result, lower values are neglected in a time series and larger values are strongly overestimated (Legates and McCabe 1999). To verify the efficiency of our models, we further estimated CC, PBIAS, and RMSE. The CC for training data set was obtained within a range of 0.92–0.98, which signifies the similarity between our observed and predicted training set. According to CC values, most well-predicted parameter was $PM_{2.5}$ and the relatively least well-predicted parameter was $O_3$. The low PBIAS (− 0.01–0.02) range indicate that model is well incorporating the bias factors also. Similarly, low RMSE value range (1.50–2.91) also asserts that our model training was efficient. As per RMSE values, CO was most efficiently predicted followed by $PM_{2.5}$, $NO_x$, and $O_3$.

## Testing results

We selected best fit models according to the training results as shown in Tables 3 and 4. For testing data sets, the range of NSE value was 0.86–0.96. Best NSE value was obtained during the prediction of $PM_{2.5}$ concentrations and least during $O_3$ predictions. The coefficient of correlation followed similar behavior, and it was observed that best CC value (0.97) was obtained for the case of $PM_{2.5}$ and least (0.86) during $O_3$ predictions. Similar to the training case, a very low range (− 0.06–0.03) of PBIAS was obtained. In addition, RMSE value range lied in the range 1.42–2.71, which is at par with the training range of 1.50–2.91. Hence, it can be established by the overall performance of models for both training and testing sets that LSTM approach is quite efficient in capturing most of the aspects of air quality prediction, which can also be visualized in the scatter plots between observed and simulated data for both training and testing cases Fig. 3.

## Comparison with previous study

As discussed, the deep learning approach has been widely used in the air quality classification and forecasting problems. For Indian context also, researchers have employed various conventional machine learning and deep learning methods for predicting the air pollutants. We have already discussed the air quality scenario of Indian capital Delhi, which makes it one of the worst polluted cities of the world. Our results show that the LSTM approach is profoundly efficient in predicting the air pollution parameters. To verify the significance of our approach, we compared our results with a couple of previous

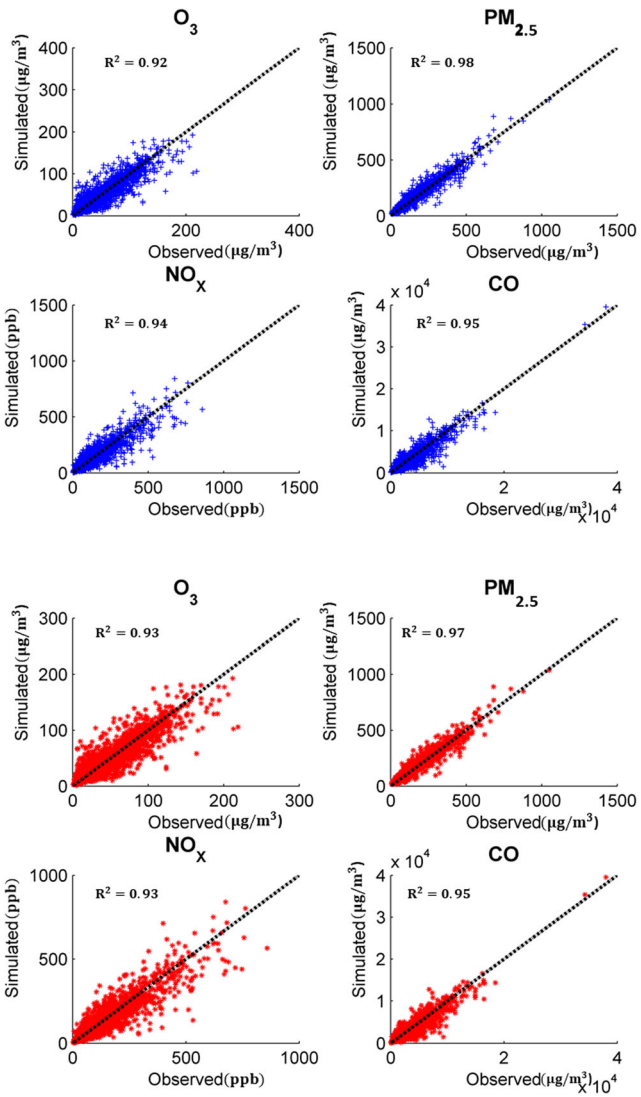


**Fig. 3** The plots for observed and simulated training (blue dots) and testing (red dots) data sets for $O_3$, $PM_{2.5}$, $NO_x$, and CO for best fit model

studies (Sekar et al. 2016a, b) carried out for the same region using other machine learning methods such as ANNs and model trees. Both the studies were aimed at predicting $O_3$, $PM_{2.5}$, $NO_x$, and CO concentrations at the same ITO intersection location. For CO and $PM_{2.5}$, it was concluded that M5P algorithm performed better than REPTree and ANNs in precisely understanding the relationships among pollutant concentrations and predictor variables. The ANN models employed in prediction of $PM_{2.5}$ were poorly performing, and the estimated CC values were in a range of 0.47–0.78. The ANN models estimate for CC lied in the range of 0.33–0.61 for CO. Similarly, the REPTree model outputs produced a CC range of 0.54–0.84 for CO, which is not competent for such predictions. Further, the correlation coefficients for M5P models, which were concluded as best for both $PM_{2.5}$ and CO predictions, were estimated to be in the range of 0.70–0.89 and 0.43–0.65. In addition, we also performed the regression analysis using the support vector machine (SVM) and relevance vector machine (RVM). The model performance of LSTM gives better RMSE values ranging 1.50–2.91 in predicting $O_3$, $PM_{2.5}$, $NO_x$, and CO concentrations, while the RMSE values of SVM and RVM models range from 14 to 833 during the simulation. Our analysis using LSTM-based approach produced the CC values of 0.97 and 0.95 for the best fit models which proves the supremacy of LSTM models over M5P, REPTree, and ANN. In addition to this, in other works, the performance of M5P, REPTree, and ANN models were compared for predicting $O_3$ and $NO_x$ using same set of input combinations. The CC obtained for ANN models varied in the range of 0.58–0.76 and 0.52–0.72 for $O_3$ and $NO_x$ respectively. Further, CC estimates for the REPTree models were observed to be in the limit of 0.51–0.76 for $O_3$ and 0.60–0.79 for $NO_x$. M5P model was decided to be the best for predicting the two parameters, which yielded a correlation coefficient value within a range of 0.60–0.82 and 0.69–0.83. We compared these values with LSTM model outputs and found that LSTM models are better predicting for both $O_3$ and $NO_x$. We obtained the CC values as 0.93 for both $O_3$ and $NO_x$, which is better as compared to ANN, M5P, and REPTree estimates. This comparison clearly indicates that long-term dependency in modelling air quality parameters could not be effectively captured by the employed models and LSTM is able to tackle this complexity.

The limitations of the present study are as follows: (i) the length of data is limited, and with the availability of recent air quality data, the validation might have improved and the efficiency of the model in projecting the future air quality could have achieved; (ii) in this study, we have considered only one station as per the availability of data and hence the ability of LSTM is not analyzed in capturing the spatial correlation. These limitations provide a direction to pursue the study in the future.

## Conclusions

This study illustrates the findings of an analysis aimed at establishing the use of LSTM approach for the prediction of hourly air pollutant concentrations (i.e., $O_3$, $PM_{2.5}$, $NO_x$, and CO). We employed latest LSTM models for our analysis. The efficiency of this approach is verified through the investigation of traffic, meteorological, and emission data obtained for a region in Delhi-NCT. Different models were formulated by combining five different combinations of input parameters. The outcomes from the present investigation are as follows:

- We found that meteorological factors play most vital role in the prediction of CO concentration, whereas all factors, including meteorological, traffic, and emission characteristics, are significant in governing the prediction of $PM_{2.5}$ and $O_3$.
- Similarly, both the meteorological data and traffic data are most important variables in $NO_x$ predictions.
- The NSE value range (0.86–0.94) obtained in our analysis proves that LSTM models used in this study are well suited for prediction of air pollutant concentrations.
- The CC range of 0.93–0.97 further confirms this fact. Lastly, we compared our results with other such studies and found that LSTM models were the most superior ones.

**Acknowledgements** We thank the editor and two anonymous reviewers for their insightful and constructive comments to improve the manuscript significantly.

## References

Ahn J, Shin D, Kim K, Yang J (2017) Indoor air quality analysis using deep learning with sensor data. Sensors 17:2476

Almaraz M, Bai E, Wang C, Trousdell J, Conley S, Faloona I, Houlton BZ (2018) Agriculture is a major source of NOx pollution in California. Sci Adv 4:eaao3477. https://doi.org/10.1126/sciadv.aao3477

Athanasiadis IN, Kaburlasos VG, Mitkas PA, Petridis V (2003) Applying machine learning techniques on air quality data for real-time decision support. In: First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland. Citeseer

Automotive Research Association of India (2007) Air quality monitoring project-Indian clean air programme (ICAP). Draft Rep. on emission factor development for Indian vehicles, Pune

Bao W, Yue J, Rao Y (2017) A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS One 12:e0180944

Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5:157–166

Briggs DJ, de Hoogh C, Gulliver J, Wills J, Elliott P, Kingham S, Smallbone K (2000) A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Sci Total Environ 253:151–167. https://doi.org/10.1016/S0048-9697(00)00429-0

Chugh S, Kumar P, Muralidharan M, et al (2012) Development of Delhi driving cycle: a tool for realistic assessment of exhaust emissions from passenger cars in Delhi. SAE Technical Paper

Fan J, Li Q, Hou J et al (2017) A spatiotemporal prediction framework for air pollution based on deep RNN. ISPRS Ann Photogramm Remote Sens Spat Inf Sci 4:15

Fu M, Wang W, Le Z, Khorram MS (2015) Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. Neural Comput & Applic 26: 1789–1797

Ghasemi A, Amanollahi J (2019) Integration of ANFIS model and forward selection method for air quality forecasting. Air Qual Atmos Health 12:59–72. https://doi.org/10.1007/s11869-018-0630-0

Gokhale S, Pandian S (2007) A semi-empirical box modeling approach for predicting the carbon monoxide concentrations at an urban traffic intersection. Atmos Environ 41:7940–7950

Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on. IEEE, pp 6645–6649

Gurjar BR, Ravindra K, Nagpure AS (2016) Air pollution trends over Indian megacities and their local-to-global implications. Atmos Environ 142:475–495

Gurjar BR, Van Aardenne JA, Lelieveld J, Mohan M (2004) Emission estimates and trends (1990–2000) for megacity Delhi and implications. Atmos Environ 38:5663–5681

Guttikunda SK, Calori G (2013) A GIS based emissions inventory at 1 km × 1 km spatial resolution for air pollution analysis in Delhi, India. Atmos Environ 67:101–111

Guttikunda SK, Goel R, Pant P (2014) Nature of air pollution, emission sources, and management in the Indian cities. Atmos Environ 95: 501–510. https://doi.org/10.1016/j.atmosenv.2014.07.006

Guttikunda SK, Gurjar BR (2012) Role of meteorology in seasonality of air pollution in megacity Delhi, India. Environ Monit Assess 184: 3199–3211

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780

PressTrust of India (2018) 40 pc of India's population likely to reside in cities by 2030: Puri. Press Trust India, India Today

Jain S, Khare M (2010) Adaptive neuro-fuzzy modeling for prediction of ambient CO concentration at urban intersections and roadways. Air Qual Atmos Health 3:203–212. https://doi.org/10.1007/s11869-010-0073-8

Kalapanidas E, Avouris N (2001) Short-term air quality prediction using a case-based classifier. Environ Model Softw 16:263–272

Kim MH, Kim YS, Lim J, Kim JT, Sung SW, Yoo CK (2010) Data-driven prediction model of indoor air quality in an underground space. Korean J Chem Eng 27:1675–1680

Kurt A, Oktay AB (2010) Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst Appl 37:7986–7992

Legates DR, McCabe GJ Jr (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 35:233–241

Li X, Peng L, Yao X, Cui S, Hu Y, You C, Chi T (2017) Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ Pollut 231:997–1004

Mallet V, Sportisse B (2008) Air quality modeling: from deterministic to stochastic approaches. Comput Math Appl 55:2329–2337

Mayer H, Gomez F, Wierstra D, Nagy I, Knoll A, Schmidhuber J (2008) A system for robotic heart surgery that learns to tie knots using recurrent neural networks. Adv Robot 22:1521–1537

Mikolov T, Joulin A, Chopra S, Mathieu M, Ranzato MA (2014) Learning longer memory in recurrent neural networks. arXiv preprint arXiv:1412.7753

Milionis AE, Davies TD (1994) Regression and stochastic models for air pollution—I. Review, comments and suggestions. Atmos Environ 28:2801–2810

Ni XY, Huang H, Du WP (2017) Relevance analysis and short-term prediction of PM2. 5 concentrations in Beijing based on multi-source data. Atmos Environ 150:146–161

Pardo E, Malpica N (2017) Air quality forecasting in Madrid using long short-term memory networks. In: International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer, pp 232–239

PTI (2018). 40 pc of India's population likely to reside in cities by 2030: Puri. Press Trust India, India Today.

Schnelle KB, Dey PR (2000) Atmospheric dispersion modeling compliance guide. McGraw-Hill, New York

Sekar C, Gurjar BR, Ojha CSP, Goyal MK (2016a) Potential assessment of neural network and decision tree algorithms for forecasting ambient PM2.5 and CO concentrations: case study. J Hazard Toxic Radioact Waste 20:A5015001. https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000276

Sekar C, Ojha CSP, Gurjar BR, Goyal MK (2016b) Modeling and prediction of hourly ambient ozone (O3) and oxides of nitrogen (NOx) concentrations using artificial neural network and decision tree algorithms for an urban intersection in India. J Hazard Toxic Radioact Waste 20:A4015001. https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000270

Sønderby SK, Sønderby CK, Nielsen H, Winther O (2015) Convolutional LSTM networks for subcellular localization of proteins. In: International Conference on Algorithms for Computational Biology. Springer, pp 68–80

Srivastava A, Jain VK (2005) A study to characterize the influence of outdoor SPM and associated metals on indoor environment in Delhi. J Environ Sci Eng 47:222–231

UN (2018) 2018 revision of world urbanization prospects. https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html. Accessed 14 April 2019

West JJ, Naik V, Horowitz LW, Fiore AM (2009) Effect of regional precursor emission controls on long-range ozone transport—part 1: short-term changes in ozone air quality. Atmos Chem Phys 9: 6077–6093

WHO, 2018. Global Ambient Air Quality Database (update 2018). World Health Orgination.

Zhang J, Zhu Y, Zhang X, Ye M, Yang J (2018) Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas. J Hydrol 561:918–929

Zhong W, Yu H, Song L, Zhang X (2011) Combined pretreatment with white-rot fungus and alkali at near room-temperature for improving saccharification of corn stalks. BioResources 6:3440–3451