Comparative study on the effects of meteorological and pollutant parameters on ANN modelling for prediction of SO₂

Reshma Unnikrishnan¹ · G. Madhu¹

© Springer Nature Switzerland AG 2019

Abstract

Variations in meteorological parameters, different transportation mechanisms, complex reaction mechanisms and insufficient control measures have made control of ambient air pollution, a challenge. Prediction mechanisms for pollutant concentrations in advance have become necessary to regulate parameters within the acceptable limits. Artificial neural network (ANN) modelling can be used for predicting the concentrations of pollutants by establishing functional relationships between complex and nonlinear predictor variables and outputs. The application of time-series ANN models to predict air quality parameters was investigated by using six pollutant variables, five meteorological parameters and three time parameters to predict the concentration of SO₂. An industrial belt in the southern part of India was selected as the study area. Two years of input parameters and feed-forward back propagation algorithm were used to construct the ANN model. Input parameters were optimized using forward selection and backward elimination techniques. Mean squared error and coefficient of determination were used to evaluate the models. The developed model exhibited very promising performance evaluation characteristics with the best model resulting in MSE of 0.0115 and R^2 value of 0.8979. The hybrid ANN models employing input parameter optimization techniques resulted in better performance characteristic values than conventional models. Any model exhibited a minimum reduction in MSE by 9% and 3% improvement in correlation. Pollutant parameters were found to affect the ANN models when compared to the meteorological parameters. Predicted values compared to the National Ambient Air Quality Standards were as low as 55% of the maximum allowable concentration in ambient air.

Keywords Artificial neural network (ANN) · Air quality modelling · Prediction · Backpropagation · SO₂ concentration · Feedforward · Time-series

Mathematics Subject Classification 92B20

JEL Classification C15 · C45 · Q35

1 Introduction

Ambient air pollution is one of the serious problems encountered by developing countries like India. In 2018, India was ranked 177 among 180 countries in the Environmental Performance Index (EPI) [1]. The study area selected is the industrial area of Ambalamugal in the

South Indian state of Kerala. It is ranked 24th amongst the critically polluted areas (CPA) in India [2]. The major industries located in the area comprise of a petroleum refinery, a phosphatic fertiliser unit, petrochemical plant and a carbon black manufacturing industry. In the urban environment, various types of pollutants are released into the atmosphere at varying concentrations, at different

✉ G. Madhu, profmadhugopal@gmail.com | ¹Division of Safety and Fire Engineering, School of Engineering, Cochin University of Science and Technology, Kochi, Kerala 682 022, India.

heights and meteorological parameters also determine the transportation, dilution, dispersion, transformation, deposition and absorption of pollutants [3].

The main air pollutants in the atmosphere are carbon monoxide (CO), nitrogen oxides (NO_x), particulate matter (PM), sulphur dioxide (SO₂) and ozone (O₃) [4]. Exposure to air pollutants escalates the risk of contravention of respiratory diseases, such as asthma, respiratory infections and chronic obstructive pulmonary disease, in children and adults alike [5]. Considering the area of study, sulphur dioxide (SO₂) and sulphur trioxide (SO₃) are the major oxides of sulphur responsible for pollution along with sulphate-containing compounds (SO₄²⁻). They are mainly produced by the combustion of fuel containing sulphur, from refining, manufacturing processes, municipal incineration and metal extraction processes. SO₂ can cause acid rains, corrosion, damage to human, plant and animal health. These pollutants also pose heightened danger due to their synergistic effects as there is prospects of their interactions, having to exist in the same medium [6].

It is important that sufficient tools are available to predict and forecast the concentrations of these harmful pollutants well in advance. Air pollution models currently being adopted are of two types (i) Deterministic or Mechanistic models and (ii) Statistical or Data-driven models [7, 8]. The former approach considers mathematical representation of various pollutant transportation mechanisms and their reactions, chemically. Hence, they tend to be strenuous, exhaustive and computationally expensive. The latter on the other hand have the ability to identify

the relationships between the input and output variables without having to evaluate all transportation mechanisms or chemical transformations of the pollutants. Artificial Neural Network (ANN) is a statistical model and its capability of learning, training and predicting with parameters that have nonlinear relationships with each other, is being utilized in this work [9]. ANN model is designed to replicate the simple function of a biological network and is used for solving complex nonlinear functions. ANN has the capability to recognize nonlinear relationships between variables and complex patterns in data sets that may not be adequately described by simple mathematical equations [6]. ANN can model relationships without making any prior assumptions regarding the data distributions [10]. If sufficiently trained, statistical models like ANN are found to be more appropriate in determining dependencies between pollutant concentration and predictor parameters than deterministic models [11].

ANN models (Fig. 1) are similar to working biological systems which consists of simple but numerous nerve cells called “Neurons” which are interconnected to form networks. Generally, there are three layers present in ANN model. The first layer is called as the input layer and the total number of nodes in this layer equal to the number of input parameters. The final layer is called as the output layer and the number of nodes equals the number of outputs. The hidden layer connects the input and output layers. Input data are introduced to the network through the input layer and the output layer provides the response of the network [12]. Neurons have

Fig. 1 Artificial neuron model

the ability to work in parallel and learn. Hence, a neural network need not be necessarily programmed, but they can learn from training samples or by means of encouragement. The resulting successful network has a high degree of error tolerance against noisy data. ANN models are designed to mimic these capabilities of biological neural systems.

The activation functions of the input and output layers can be either linear or nonlinear, while the hidden layer can have only nonlinear activation functions [13]. Sigmoid functions are widely used as there exist simple relationship between the value of the function and derivative value at any point. This reduces the computational complexity of the network [14]. The hyperbolic tangent and logarithmic sigmoid functions are defined in Eqs. (1) and (2), respectively [15]:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (1)$$

$$f(x) = \ln \left(\frac{1}{1 + e^{-x}} \right) \quad (2)$$

The most common network used for prediction modelling is the multilayer perceptron feed-forward network where the information flows in one direction from input to output nodes [16]. The study was to develop an optimized feed-forward ANN model employing back propagation algorithm suitable for predicting the air quality parameters for the South Indian industrial area of Ambalamugal. A multilayer feed-forward ANN model is characterized by one input, at least one hidden and one output layers, each composed of neurons [13]. ANN models with Levenberg–Marquardt backpropagation algorithm having one hidden layer is used for prediction models usually [12, 15, 17]. Backpropagation algorithm feeds back the error produced by neural networks to the nodes to modify the connection weights and thresholds [12]. For this particular study intended at predicting the air quality parameters, time-series modelling was utilized. Time-series modelling analyses the data which is already available in the past and by extrapolating the future values based on this historical data [18]. It employed a type of dynamic filtering, i.e. the parameter values in the past of one or more time series were utilized to predict the data in future. Referred to as dynamic neural networks, they utilized tapped delay lines used for nonlinear filtering and prediction. ANN models have been effectively used to determine nonlinear solutions to prediction problems. ANN models can be effectively used to approximate any quantifiable function if they are designed and trained by appropriate historical data. ANN has the ability to predict values even in the presence of noise data and it learns and determines the patterns and generalizations in the available data.

If seasonal variations are duly considered, a full year of data is sufficient for developing statistical models if the emissions are present throughout the year [19]. The data sets have to be partitioned into training, validation and testing data sets such that every element of a subset represents the entire data set [15]. The data sets used in the modelling have to be cleaned, randomized and normalized. Removal of unusual values, errors and outliers is necessary before modelling as the network learns according to the input data sets. If introduced to an erroneous data set, the network will learn accordingly and ultimately results in erroneous results. The data sets have to be normalized to ensure that all the input values fall into comparable range. If not normalized, inputs having higher numerical magnitude have a tendency to mask the inputs which are of numerically smaller values [16].

As the number of input parameters increases, the complexity of the developed ANN model also increases which results in degraded model performance. Hence it is imperative that input optimization techniques are applied to narrow down the parameters to optimum by suitably eliminating those variables that have the least effect on the model's efficiency [15]. Forward selection starts with no predictor variable in the model and subsequently adds those variables that have the highest correlation with the target output. Backward elimination begins with all predictor input variables and subsequently eliminates those parameters that provides the least increase in the squared error [15].

The number of hidden layers, the number of neurons constituting the hidden layers, the transfer functions and the training algorithm also affect the performance of the ANN models [13]. Training is the process of determining connection weights and thresholds to minimize the difference between the actual and predicted outputs [12]. Mean squared error (MSE) and coefficient of determination (R^2) are some of the parameters used for evaluation of developed networks. The higher the value of coefficient of determination ($R^2 = 1$) and lower the value of mean squared error ($MSE = 0$), the network's performance is considered to be superior [6]. The work also optimizes input parameters based on the performance of the ANN model using forward selection and backward elimination. Comparison of performance of optimized models is achieved using model performance evaluation. The effects of meteorological and pollutant parameters on the prediction capability of the ANN model are also compared.

There is a crucial need to monitor the concentrations of various pollutants and the relationship of various meteorological and pollutant parameters in the prediction of ambient air quality parameters. Even though studies are available for prediction of ambient air quality parameters, air quality models for monitoring pollutant concentrations

are not popular in developing world. In particular, in critically polluted area like Ambalamugal in Kerala where the study is being conducted, even though continuous ambient air quality monitoring facilities are available, additional tools like ANN model is to be used to predict the concentrations of air quality parameters considering the equipment is taken out for maintenance or calibration. During the severe floods which affected the state of Kerala recently on 8 August 2018 due to unusually high rainfall during the monsoon season, the monitoring equipment was found not being functional. It is imperative that auxiliary mechanisms or techniques that can predict the air quality parameters are evolved, in addition to direct measurement devices installed in these areas. Additionally, comparison of effects of pollutant and meteorological parameters on the prediction capabilities of ANN-based modelling is to be conducted. This study aims to determine the capability of ANN networks in prediction of air quality parameters in an industrial area and the same is demonstrated by assessing the prediction capability of concentration of sulphur dioxide (SO_2) using various input parameters. Owing to site specificity, ANN models are individually constructed for each location. A neural network constructed and trained to perform for a particular monitoring location will not be effective in predicting the pollutant concentrations at a different location as the predictor variables would vary with sites [11, 17]. Hence the boundary of the applicability of the developed model is limited to the site under study.

2 Materials and methods

Artificial Neural Networks are statistical models used for training and prediction of outputs that have nonlinear relationships with their corresponding inputs. The topology, ANN parameters and the learning algorithm have been selected satisfactorily to predict the outputs with required performance level.

2.1 Study area

The study area selected was Ambalamugal, an industrial suburb in the South Indian state of Kerala. It is located 14 km towards East from Ernakulam in Kerala and has an altitude of 12 m above the sea level.

2.2 Software

The Neural Network Toolbox of MATLAB R2018a (The MathWorks Inc. USA) was used for constructing the air quality prediction model. The Neural Network Toolbox

allowed selection of various parameters for configuring the desired network architecture.

2.3 Data pre-processing

The prediction model was developed using data sets collected from September 2016 to September 2018 stretching over a period of 2 years. Pollutant data was collected from the three continuous ambient air quality monitoring stations maintained by the petroleum refinery located in the study area. The facility is capable of measuring six air quality parameters, namely sulphur dioxide (SO_2), nitrogen oxide (NO_x), ammonia (NH_3), carbon monoxide (CO), particulate matter of size less than $10\text{ }\mu\text{m}$ in diameter (PM_{10}) and particulate matter of size less than $2.5\text{ }\mu\text{m}$ in diameter ($\text{PM}_{2.5}$). Meteorological data were collected from Indian Meteorological Department (IMD). Meteorological parameters being considered were surface temperature (T), rainfall (RF), relative humidity (RH), wind direction (WD) and wind velocity (WV). The time parameters considered for the study were month of the year (MY), day of the month (DM) and day of the week (DW). This was achieved by considering values from 1 to 12 for January to December; values from 1 to 31 for the corresponding day of the month; and values 1–7 for the corresponding day of the week (1—Sunday to 7—Saturday).

The collected data were pre-processed by cleaning, normalizing and randomizing the input data sets. Any unclear data, anomalies, outliers or errors were removed from the data sets. The original inputs were normalized for convergence of the model using Eq. (3), and normalized outputs were converted back to original data sets using Eq. (4), given below:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \times (r_{\max} - r_{\min}) + r_{\min} \quad (3)$$

$$X = \left(\left(\frac{X_{\text{norm}} - r_{\min}}{r_{\max} - r_{\min}} \right) \times (X_{\max} - X_{\min}) \right) + X_{\min} \quad (4)$$

where X_{norm} was the normalized value, X was the original value, $X_{\min}$ and $X_{\max}$ were the minimum, and maximum values of X , $r_{\min}$ and $r_{\max}$ were the values of -0.9 and 0.9 , respectively.

2.4 ANN model

Out of the total data sets collected, 70% of the data sets was used for training the constructed neural network. 15% of the data sets was used for validation and the rest for testing purpose. In this study, a feedforward time-series neural network was used. The model comprised of three distinct layers—input layer, hidden layer and output layer.

Generally, one hidden layer was considered for the construction of ANN models used for prediction problems [14]. The number of nodes in the input layer equalled the number of input parameters, which is 14. The number of neurons in the hidden layer was decided by observing the performance of the network during trial and error process using varying number of neurons. The number of hidden neurons was determined by considering up to $(2N + 1)$ values of neurons; where N was the number of input parameters employed in each model [16]. The number of neurons which provided the best performance was considered as the final value. The number of nodes in the output layer equalled the number of output to be obtained from the network model. The network was trained using Levenberg–Marquardt backpropagation algorithm. The algorithm was based on the error correction gradient descent learning method. Time-series prediction models were used in the study which extrapolated the output data based on the historical input data which were already available. Here, the multilayer perceptron model used input data whose seasonal variations were retained, by prediction using time series models and the future data was predicted as output. The multilayer perceptron model performed the following functional mapping as given in Eq. (5).

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-n}) \quad (5)$$

where y_t was the estimated output, and $y_{t-1}, y_{t-2}, \dots, y_{t-n}$ were the training pattern which consisted of a fixed number (n) of lagged observations of the series [18]. The number of delays used for time-series forecasting was determined by trial and error process by observing the performance of the network using varying number of

delays, keeping all the other parameters constant. The number of delays which resulted in the best performing model was finalised as the delay value.

The report was prepared based on the pollutant, meteorological and time parameters. Accordingly, three different types of architectures were considered. Model A was constructed using all the 14 input parameters. Model B was constructed using only SO_2 , meteorological and time parameters. Here the other pollutant parameters were excluded. Model C was constructed using only the pollutant and time parameters. In Model C, the meteorological parameters were considered excluded.

Typical architecture of the ANN model A constituting of the 14 input parameters (six pollutant + five meteorological + three time parameters) and one output is provided in Fig. 2.

Typical architecture of the ANN model B constituting of the nine input parameters (five meteorological + three time parameters + SO_2), and one output is provided in Fig. 3.

Typical architecture of the ANN model C constituting of the nine input parameters (six pollutant + three time parameters) and one output is provided in Fig. 4.

Flowchart of the general methodology followed in the construction and execution of the ANN model is provided in Fig. 5.

2.5 Input parameter optimization

Model A was constructed using all the 14 input parameters; Model B using only the meteorological and time parameters and SO_2 ; and Model C using only the pollutant and time parameters. Three different models were

Fig. 2 MLP architecture for air quality prediction using all parameters

Fig. 3 MLP architecture for air quality prediction using meteorological and time parameters + SO_2

Fig. 4 MLP architecture for air quality prediction using pollutant and time parameters

considered within a particular type of model, to understand the relevance of air quality, time and meteorological parameters in the performance of the ANN model. Model 1 was constructed using all the input parameters based on the original data set. Model 2 was constructed by choosing input parameters with forward selection technique. Model

C was constructed using ANN model with backward elimination technique. The performance of the ANN network was then evaluated to conclude which input optimization technique provides satisfactory results and the network was updated based on the optimized inputs. The list of the ANN models considered in this study is provided in Table 1.

Fig. 5 Flowchart of ANN modelling method **Table 1** List of ANN models considered in the study

ANN model	Model description
A1	ANN model constructed with all 14 input parameters (six pollutant + five meteorological + three time parameters)
A2	ANN model constructed with all 14 input parameters and Forward Selection (six pollutant + five meteorological + three time parameters)
A3	ANN model ANN model constructed using all 14 input parameters and Backward Elimination (six pollutant + five meteorological + three time parameters)
A4	ANN model constructed with the input parameters after optimization (six pollutant + five meteorological + three time parameters)
B1	ANN model constructed with all nine input parameters (five meteorological + three time parameters + SO ₂)
B2	ANN model constructed with all nine input parameters and Forward Selection (five meteorological + three time parameters + SO ₂)
B3	ANN model ANN model constructed using all nine input parameters and Backward Elimination (five meteorological + three time parameters + SO ₂)
C1	ANN model constructed with all nine input parameters (six pollutant + three time parameters)
C2	ANN model constructed with all nine input parameters and Forward Selection (six pollutant + three time parameters)
C3	ANN model ANN model constructed using all nine input parameters and Backward Elimination (six pollutant + three time parameters)

2.6 Model performance evaluation criteria

Determination of the performance of the constructed network depends on the degree of convergence of the network predicted outputs and the actual output values. Model performance evaluation criteria used in the evaluation of the ANN model were Mean Squared Error (MSE) and Coefficient of determination (R^2). The equations for calculation of MSE and R^2 are given in Eqs. (6) and (7), respectively:

$$\text{MSE} = \sum_{i=1}^n \frac{(x_i - y_i)^2}{n} \quad (6)$$

$$R^2 = 1 - \frac{\sum (x_i - y_i)^2}{\sum y_i^2 - \frac{\sum y_i^2}{n}} \quad (7)$$

where x_i denoted the actual output, y_i denoted the predicted outputs and n was the total number of observations [4, 20].

3 Results and discussion

The number of hidden neurons was determined by considering up to $(2N + 1)$ values of neurons; where N is the number of input parameters employed in the model. Here,

there were 14 input parameters, ANN model developed was trained using varying number of hidden neurons from 1 to 29 ($2 \times 14 + 1 = 29$); maintaining all the remaining conditions the same. Comparing the results for 1 to 29 neurons, it was concluded that the best performance of the model was noted with 15 number of hidden neurons which resulted in MSE value of 0.002 and coefficient of determination as 0.987. Figure 6 represents the graphical variation of MSE and R^2 values with respect to the number of hidden neurons.

The number of delays was decided based on trial and error by varying the number of delays. All the other parameters within the network were maintained the same and the performance of the networks with varying number of delays were then evaluated. Comparing the results, it was found that the best performance of the model was noted with number of delays as 2. Incidentally, this was the default set value for number of delays in the MATLAB software's timer-series application. Further increase in delays was not found to improve the performance of the ANN models. The model resulted in MSE value of 0.016 and coefficient of determination as 0.863. Figure 7 represents the graphical variation of MSE and R^2 values with respect to the number of delays.

Model A was developed using 9184 input data sets (14 parameters $\times$ 656 observations) collected from September, 2016 to September, 2018 stretching over a period of 2 years. The data sets were also partitioned such that 70% of the data (6440 data sets) was used for training, 15%

Variation of Mean Squared Error (MSE) with Number of neurons

Variation of Coefficient of Determination (R^2) with Number of neurons

Fig. 6 Variation of mean squared error (MSE) and coefficient of determination (R^2) with number of neurons

Fig. 7 Variation of mean squared error (MSE) and coefficient of determination (R^2) with number of delays

(1372 data sets) for validation and 15% for testing (1372 data sets). Similarly, Model B was developed using 5904 input data sets (9 parameters $\times$ 656 observations). Model C was developed using 9 input parameters which consisted of pollutant and time variables. This prediction model B was developed using 5904 input data sets (9 parameters $\times$ 656 observations).

Table 2 provides the details of the performance evaluation of the different models developed using input optimization techniques. Models which combined ANN modelling with input optimization technique provided better

performance than conventional ANN model which considered all inputs without any optimization. There was reduction in the error value and increase in correlation, when compared to the conventional model. However, comparing the results, backward elimination technique in general was found to provide better results than forward selection.

Out of Model 'A' ANN models, Model A4 which was constructed using input parameters chosen utilizing input optimization techniques, provided better result when compared to the remaining models. In Model 'B', the ANN model combined with backward elimination provided

Table 2 Performance evaluation of ANN models

Model	Model parameters	Inputs excluded from the model	No. of iterations	MSE	R^2
A1	MY, DM, DW, SO ₂ , NO _x , NH ₃ , CO, PM ₁₀ , PM _{2.5} , T, RF, RH, WD, WS	–	11	0.0188	0.8325
A2	DM, DW, SO ₂ , NH ₃ , CO, WS	MY, NO _x , PM ₁₀ , PM _{2.5} , T, RF, RH, WD	11	0.0129	0.8679
A3	DM, DW, SO ₂ , NH ₃ , CO, PM ₁₀ , RF, WD	MY, NO _x , PM _{2.5} , T, RH, WD	11	0.0158	0.8617
A4	DM, DW, SO ₂ , NH ₃ , CO, PM ₁₀ , RH, WD, WS	MY, NO _x , PM _{2.5} , T, RH	10	0.0124	0.8727
B1	MY, DM, DW, SO ₂ , T, RF, RH, WD, WS	–	15	0.0162	0.8675
B2	MY, DW, SO ₂ , T, RF, WD	DM, RH, WS	13	0.0152	0.8914
B3	MY, DW, SO ₂ , T, RF, WD	DM, RH, WS	10	0.0124	0.8935
C1	MY, DM, DW, SO ₂ , NO _x , NH ₃ , CO, PM ₁₀ , PM _{2.5}	–	9	0.0171	0.8559
C2	MY, DW, SO ₂ , NH ₃ , PM _{2.5}	DM, NO _x , CO, PM ₁₀	14	0.0130	0.8944
C3	MY, DW, SO ₂ , NH ₃ , PM _{2.5}	DM, NO _x , CO, PM ₁₀	15	0.0115	0.8979

the best results and Model 'C' also converged in similar lines. Out of all the 10 models evaluated, the best result was exhibited by Model C3. This model was constructed using pollutant and time parameters only. According to this model, 5 parameters namely, month of the year, day of week, concentration of SO_2 , NH_3 and $\text{PM}_{2.5}$ were found to contribute maximum to the prediction capability of the model.

Figure 8 represents the regression plots of Model A1, A2, A3 and A4.

Figure 9 depicts the variation of actual and predicted SO_2 concentrations using best model C3.

Also the predicted concentration of SO_2 concentrations was compared with the National Ambient Air Quality

Standards values [21] and were all below the allowable limit of $80 \mu\text{g}/\text{m}^3$ (24 h TWA value). The maximum value of SO_2 obtained from predicted values was only 55% of the maximum allowable concentration in ambient air.

4 Conclusions

The current experimental research utilized feed-forward back propagation artificial neural network modelling to predict the air quality parameters based on 6 pollutant, 5 meteorological and 3 time parameters. The prediction capability of the model was demonstrated by predicting the concentration of sulphur dioxide in an industrial

Fig. 8 Regression plot of model A networks

Fig. 9 Variation of actual and predicted SO_2 concentrations using best model C3

area using time-series ANN modelling. Different ANN models were developed based on parameter selection techniques to determine the input parameters which affected the prediction capability of the ANN model. Similarly different ANN models were also constructed and evaluated to study the effects of pollutant and meteorological variables.

- Out of all the models, the conventional model with all parameters exhibited reduced performance which implied input optimization and subsequent parameter selection resulted in improved performance. The other models exhibited an average reduction in MSE by 26% and an average improvement in correlation by 4%.
- Comparison of performance of the best performing models of A, B and C ANN models showed that the best performance was exhibited by ANN model having pollutant and time parameters only. The minimum MSE obtained in any model was 0.0115 and maximum R^2 value of 0.8979.
- The best result was provided by models incorporating pollutant variables since the concentration of SO_2 pollutant was being predicted in the study. Additionally, any model trained and evaluated showed a minimum reduction in MSE of 9% and improvement in correlation by 3%.
- When compared with the National Ambient Air Quality Standards values, the predicted concentrations of SO_2 were found to fall well below the allowable limit; as much as 55% of the maximum allowable concentration in ambient air. The rate of gaseous SO_2 emissions into the ambient air and their control measures in place could be monitored by similar comparisons and were found to be in agreement.
- The results revealed that ANN models can be effectively utilized for predicting the concentration of pollutants. As far as prediction capacity of the ANN models for primary air pollutants was concerned, the models exhibited very promising results. However, requirements of including additional parameters, while trying to predict secondary air pollutants where other reaction mechanisms have also to be considered, has to be studied on.
- Comparing the results it was found in general that, all the models were showing values of MSE closer to zero and R^2 closer to unity. Methods of input optimization and considerations of whether to include pollutant or meteorological parameters—all helped in fine-tuning the predicted results and reduced the disparity between the actual and predicted outputs.

Also, further studies could be conducted to understand the mechanism by which ANN was evolving relationships between input and output variables. Hybrid ANN model

can be effectively utilized to predict the concentration of SO_2 and similar air quality parameters in an industrial area.

Acknowledgements The authors wish to thank Ms. BPCL-Kochi Refineries Limited, Kochi, India, and Ms. Indian Meteorological Department, Thiruvananthapuram, India, for providing pollutant and meteorological data, respectively.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human and animals rights The present research does not involve human participants and/or animals.

References

1. Yale (2018) Environmental Performance Index—global metrics for the environment: ranking country performance on high-priority environmental issues, pp 1–4
2. Kerala State Pollution Control Board (2010) Action Plan for Greater Kochi Area, Executive Summary, pp 1–89
3. Goyal SK, Chalapati Rao CV (2007) Assessment of atmospheric assimilation potential for industrial development in an urban environment: Kochi (India). *Sci Total Environ* 376:27–39. <https://doi.org/10.1016/j.scitotenv.2007.01.067>
4. Oduber F, Calvo AI, Blanco-Alegre C, Castro A, Vega-Maray AM, Valencia-Barrera RM, Fernández-González D, Fraile R (2019) Links between recent trends in airborne pollen concentration, meteorological parameters and air pollutants. *Agric For Meteorol* 264:16–26. <https://doi.org/10.1016/j.agrformet.2018.09.023>
5. Kim D, Chen Z, Zhou LF, Huang SX (2018) Air pollutants and early origins of respiratory diseases. *Chronic Dis Transl Med* 4:75–94. <https://doi.org/10.1016/j.cdtm.2018.03.003>
6. Fraile R, Monsalve F, Tomás C (2013) Influence of meteorological parameters and air pollutants onto the morbidity due to respiratory diseases in Castilla-La Mancha, Spain. *Aerosol Air Qual* 13:1297–1312. <https://doi.org/10.4209/aaqr.2012.12.0348>
7. Athira V, Geetha P, Vinayakumar R, Soman KP (2018) DeepAirNet: applying recurrent networks for air quality prediction. *Procedia Comput Sci* 132:1394–1403. <https://doi.org/10.1016/j.procs.2018.05.068>
8. Azid A, Zain SM, Latif MT, Juahir H, Osman MR (2013) Feed-forward artificial neural network model for air pollutant index prediction in the southern region of Peninsular Malaysia. *J Environ Prot (Irvine, Calif)* 4:1–10. <https://doi.org/10.4236/jep.2013.412a1001>
9. Samarasinghe S (2010) Neural networks for nonlinear pattern recognition. In: *Neural networks for applied sciences and engineering*. <https://doi.org/10.1201/9781420013061.ch3>
10. Ghazali S, Ismail LH (2013) Air quality prediction using artificial neural network. *Int J Adv Comput Eng Appl* 3:5:1–5
11. Feng X, Li Q, Zhu Y, Hou J, Jin L, Wang J (2015) Artificial neural networks forecasting of PM_{2.5} pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos Environ* 107:118–128. <https://doi.org/10.1016/j.atmosenv.2015.02.030>
12. Md Kamal M, Jailani R, Shauri RL A (2006) Prediction of ambient air quality based on neural network technique. In: *SCORED 2006—proceedings on 2006 4th student conference on research and development. Towards enhancing res. excell.*

- reg., pp 115–119, 27–28. <https://doi.org/10.1109/SCORE.D.2006.4339321>
13. George J, Arun P, Muraleedharan C (2018) Assessment of producer gas composition in air gasification of biomass using artificial neural network model. *Int J Hydrogen Energy* 43:9558–9568. <https://doi.org/10.1016/j.ijhydene.2018.04.007>
 14. Fausett L (1994) Fundamentals of neural networks—architectures, algorithms and applications. Pearson Publications, London, pp 21–44
 15. Cabaneros SMLS, Calautit JKS, Hughes BR (2017) Hybrid artificial neural network models for effective prediction and mitigation of urban roadside NO₂ pollution. *Energy Procedia* 142:3524–3530. <https://doi.org/10.1016/j.egypro.2017.12.240>
 16. Shakerkhatibi M, Mohammadi N, Benis KZ, Sarand AB, Fatehifar E, Hashemi A (2015) Using ANN and EPR models to predict carbon monoxide concentrations in urban area of Tabriz. *Environ Health Eng Manag J* 2:117–122
 17. Elangasinghe MA, Singhal N, Dirks KN, Salmond JA (2014) Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmos Pollut Res* 5:696–708. <https://doi.org/10.5094/apr.2014.079>
 18. Benkachcha S, Benhra J, El Hassani H (2015) Seasonal time series forecasting models based on artificial neural network. *Int J Comput Appl* 116:0975–8887
 19. Carslaw DC, Carslaw N (2007) Detecting and characterising small changes in urban nitrogen dioxide concentrations. *Atmos Environ* 41:4723–4733. <https://doi.org/10.1016/j.atmosenv.2007.03.034>
 20. Shen J, Chen J, Zhang X, Zou S, Gao Z (2017) Outdoor and indoor ozone concentration estimation based on artificial neural network and single zone mass balance model. *Procedia Eng* 205:1835–1842. <https://doi.org/10.1016/j.proeng.2017.10.253>
 21. National Ambient Air Quality Standards (2009) Central Pollution Control Board Notification in the Gazette of India, Extraordinary, New Delhi

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.