Multi-horizon forecasting of average daily $PM_{2.5}$ concentration: A case study of air pollution in Singapore

Anjali Tripathi[1], Manoj Hari[1], Vinay Anand Tikkiwal[3], Bhishma Tyagi[1], Arun Kumar[2]

[1]*Department of Earth and Atmospheric Sciences*
[2]*Department of Computer Science and Engineering*
National Institute of Technology Rourkela
Odisha, India
[3]*Department of Electronics and Communication Engineering*
Jaypee Institute of Information Technology
Noida, India
tripathianjali331@gmail.com, manoj_h@nitrkl.ac.in, vinay.anand@jiit.ac.in,
tyagib@nitrkl.ac.in, kumararun@nitrkl.ac.in

***Abstract*—Understanding the spatial and temporal dynamics of surface concentrations of particulate matter ($PM_{2.5}$ and PM_{10}) is essential in air quality modelling and climate research. Forecasting of air pollutants is necessary to understand the variation of pollutants and to plan and implement air pollution control measures. This work proposes a methodology for forecasting $PM_{2.5}$ concentration using various meteorological parameters over multiple time horizons. The proposed deep learning-based models forecast the daily average values concentration for $PM_{2.5}$ in Singapore. The performance metrics indicate the efficacy of the proposed model in forecasting $PM_{2.5}$ concentrations over different horizons. The analysis shows that $PM_{2.5}$ concentrations are best forecasted for the 3-days ahead scenario, with RMSE and MAPE being 3.767 µg/m3 and 6.82%, respectively.**

Keywords—LSTM, air quality, machine learning, ERA5, aerosol

I. INTRODUCTION

Air quality refers to air contamination by various pollutants like dust, smoke, smog, and air impurities. The leading cause of air pollutants and poor air quality are anthropogenic activities, soaring urbanization, and thriving industrialization. Air pollution is among the top five global mortality risk factors [1]. In 2012, World Health Organisation allocated one out of nine deaths to air pollution [2], among them, three million deaths are only due to outdoor air pollution. Pollution due to Particulate Matter (PM) is considered the third ultimate reason of death for 2017 in India, including 56% because of vulnerability to the outdoor $PM_{2.5}$ accumulation and 44% allocated for domiciliary air pollution [3]. Urban areas like Delhi, Kolkata, and Mumbai, with a population exceeding 46 million, face seriously deteriorated air quality due to massive-scale expansion in anthropogenic activities and population density [4]. The long-term effects of air pollution are hazardous diseases like chronic asthma, cardiovascular diseases, pulmonary insufficiency, cardiovascular mortality, and perinatal disorders, leading to infant mortality in adult age [5]. A similar problem persists in urbanized areas, which need immediate attention and better prediction to reduce the burden of diseases and improve the quality of life of their residents. The need to regulate air quality is vital in places with high urbanization and smaller area, e.g., Singapore, for a better health perspective of the residents and visitors.

The air quality of Singapore is compromised due to numerous pollutant sources like forest fires, industries, and motor vehicles. The pollutants in the atmosphere include Ozone (O_3), Particulate matter ($PM_{2.5}$ and PM_{10}), Nitrogen dioxide (NO_2), etc. Particulate matters are classified according to their size, composition and source of origin. Understanding $PM_{2.5}$ temporal and spatial characteristics demand rigorous in-situ networks. Integrating the in-situ observations with sophisticated machine learning tools helps project the variations and trends. Artificial Intelligence (AI) proposes modern logical approaches capable of modelling the complicated and abridged character of the occupant's interactions with their thermal environment [6]. Machine Learning (ML) subset of AI, and thus, ML models make it easy to make predictions for large and complex data with more accuracy and better results; especially financial time series prediction using LSTM model effectively captures the long-term dependencies in the time series [7]. Many studies successfully employed the ML technique to predict $PM_{2.5}$. Some noteworthy ML methods include neural networks, Random Forest Method, and the Ensemble method. ML models to forecast air quality are popular and extensively explored globally. A case study on air quality in Madrid predicted using LSTM depicted its efficiency in air quality prediction compared to other models [8]. Hussain et al. [9] make air quality predictions using the traditional k-NN and non-traditional Long Short-Term Memory (LSTM) model. The output power of the Photovoltaic system is forecasted by [10] using the LSTM model for hourly datasets of a year, and the authors concluded that LSTM performs better than other models. The number of covid cases is forecasted by [11] using the hybrid GWO-LSTM model, and the results are compared with the ARIMA model; again, the hybrid model is found to give better results. The Multi-step forecasting of Global solar radiation of arid zones is performed by [12] using the LSTM model, and common statistical errors are observed. The authors [13] predicted the Stock Market Price using the LSTM model, having multiple inputs and outputs. Also, the LSTM model is used for Gold price forecasting [14] along with additional convolutional layers. The authors in [15] studied the Convolutional LSTM model for Spatiotemporal correlation, and it is found that the dual-stage attention mechanism can easily abolish extraneous information. The author in [16] proposed a hybrid random forest with a linear model (HRFLM) for predicting cardiovascular disease. Also, techniques like the Random Forest method, Support vector machine (SVM), Decision trees, k-NN, etc., are used for the

prediction. Sarkar et al. [17] presented ML techniques, like SVM, ANN, GA, etc., to predict occupational accidents. Air pollution forecasting for pollutants like $PM_{2.5}$ and various harmful gases in Taiwan is done by Y.T. Tsai et al. [18] using LSTM and Recurrent Neural Network techniques (RNN). R. Casado-Vara et al. [19] developed a deep-learning-based platform to forecast the online web traffic of a web server with an LSTM network. M.L. Shen et al. [20] propose a study using the deep learning method to predict multinational trade data.

The air quality of Singapore is explored using various in-situ observations and numerical models employed for its prediction. However, ML techniques are yet to mark their importance in air quality prediction over a region like Singapore. The LSTM network is one of the most prior recurrent networks and is efficient in most time series learning problems [21]. The present work uses an LSTM model for the multi-horizon forecast of $PM_{2.5}$ in Singapore. The input data consists of meteorological parameters like Temperature, Pressure, Relative Humidity, and wind speed, with PM_{10} and $PM_{2.5}$.

II. Dataset Description

Singapore city-state is situated on the southern tip of the Malay Peninsula (1° 17' 24.97" N and 103° 51' 7.05" E), consisting of Singapore Island in the shape of a diamond and some small islets. Singapore experiences a hot and humid climate throughout the year, with rainfall occurring almost every month [22]. This study selects in-situ sites that spatially cover Singapore's air quality dynamics, as shown in Fig. 1.

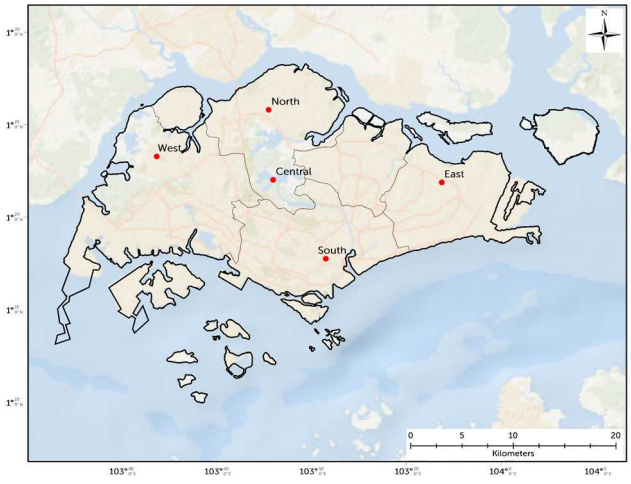

Fig. 1. Site map of Singapore with corresponding in-situ/analysis sites (red marked) used in the study.

In Singapore, the National Environmental Agency (NEA) monitors the air quality and is responsible for daily measures of ambient air pollutants $PM_{2.5}$ and PM_{10}. This study focuses on predicting $PM_{2.5}$ using the LSTM model for the year 2019 to bring out the model potential in prediction. Table I shows the maximum and minimum ranges of the variables. The correlation between the parameters $PM_{2.5}$, PM_{10}, Wind Speed (WS), Temperature (T), Relative Humidity (RH), and Pressure (P) is shown in Table II and Table III for north and south sites, respectively [23].

It can be observed in Table I that the range of $PM_{2.5}$ value is different from Relative Humidity (RH) and Wind Speed (WS); hence, $PM_{2.5}$ is found to be poorly correlated with RH and WS, as shown in Table II and III.

TABLE I. Statistical Variables and Values Range at the North and South Sites

Parameters	Value Ranges	
	North	South
$PM_{2.5}$ ($\mu g\ m^{-3}$)	19 – 233	15 – 267
PM_{10} ($\mu g\ m^{-3}$)	11 – 165	10 – 158
T (°C)	23 - 29	23 – 29
P (hPa)	1002 – 1011	1002 – 1-11
RH (%)	60 – 94	60 – 94
WS (ms^{-1})	0 – 6	0 – 6

While $PM_{2.5}$ is highly correlated with PM_{10} from value ranges, it is evident that $PM_{2.5}$ and PM_{10} lie around the same ranges. While $PM_{2.5}$ is poorly correlated with RH, as observed from the correlation matrix.

TABLE II. Correlation Matrix of Variables for the North Site

North	PM_{10}	T	P	RH	WS	$PM_{2.5}$
PM_{10}	1					
T	0.07	1				
P	0.06	–0.32	1			
RH	–0.06	–0.34	–0.11	1		
WS	–0.08	–0.12	0.29	–0.41	1	
$PM_{2.5}$	0.88	0.07	0.03	–0.07	–0.11	1

TABLE III. Correlation Matrix of Variables for the South Site

South	PM_{10}	T	P	RH	WS	$PM_{2.5}$
PM_{10}	1					
T	0.05	1				
P	0.13	–0.32	1			
RH	–0.12	–0.34	–0.11	1		
WS	0.06	–0.12	0.29	–0.41	1	
$PM_{2.5}$	0.84	0.06	0.09	–0.09	–0.05	1

Fig. 2 represents the variation of data that is used for LSTM modelling. For improved model predictions, in addition to the primary predictor ($PM_{2.5}$), the analysis is enriched by incorporating basic meteorological conditions that impact the ambient air quality.

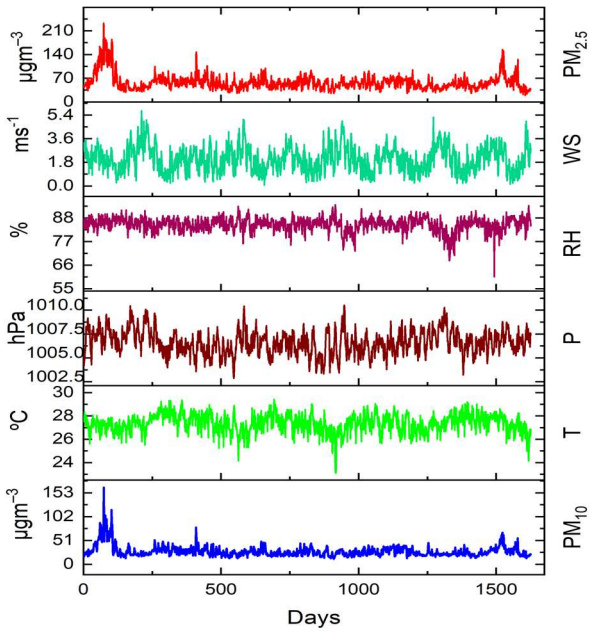

Fig. 2. Input data used for the model.

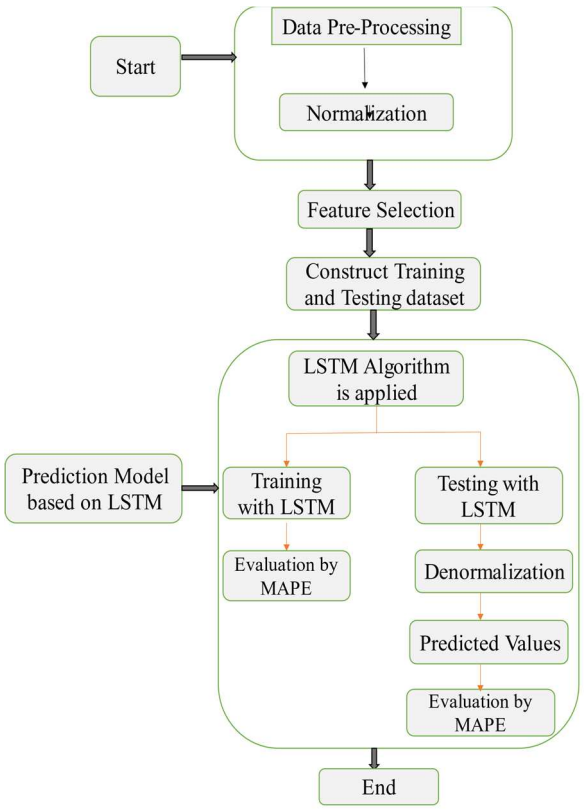

Fig. 3. Flowchart representing the processes adapted for the prediction using LSTM.

The meteorological data: temperature (T, ºC), Pressure (P, hPa), relative humidity (RH, %) and wind speed (WS, m s^{-1}) is extracted from ECMWF Reanalysis 5th Generation (ERA5) bounded to its site location.

III. Methodology Formulation

Through the meteorological sites of Singapore, informative data is obtained for the north and south sites of Singapore. Then, the LSTM model is applied, which is divided in three steps. The flowchart of the model is shown in Fig.3. The steps involved are as follows in the subsection.

A. Data Pre-processing

Firstly, a dataset is acquired with the help of data collected from different sites and then combined in a proper format. Then, python libraries like NumPy, Pandas, and Matplotlib are imported to handle the data accordingly. From the obtained dataset, missing values are identified and handled, including finding the mean for missing values or deleting some particular row/column as per requirement. For Machine learning models, it is necessary to encode the categorical data; hence the data is encoded. Then, the entire dataset is split into training and testing datasets in the ratio of 80:20. Lastly, feature scaling is done to put the values in a specific range; in this model, normalization is used for feature scaling.

B. Formation of LSTM Model

The LSTM model is used for forecasting, which stands for Long Short-Term Memory from deep learning [24]. LSTM comes under various recurrent neural networks (RNNs), which tend to learn long-term dependencies such as sequence prediction problems. An LSTM network comprised of input, hidden and output layers. Also, it consists of the cell which helps in remembering the past, which consists of three gates: (a) input gate, (b) forget gate and (c) output gate. The functions of these three gates:

(a) The input gate declares the part of the new input which can be appended to the cell state.

(b) The forget gate determines which part of the cell state should be removed depending on the hidden state, current input and forget gate weights.

(c) Based on output gate weights, hidden state, and current input, the output gate determines how current output gets affected due to cell state [25].

Weights are essential in helping the model which part of the cell to be forgotten and remembered. It helps the LSTM model to successfully detain long-term dependencies in their cell and solve the problem of vanishing gradients that the traditional RNN models face.

The hidden layer holds memory cells with input and output gates, and these gates are sigmoid activation functions where the output value ranges between 0 and 1. An output of 0 indicates that the gates block every information, while 1 means that the gates allow every piece of information [26].

The structure of the memory block is shown in Fig. 4. below, and the memory block is a fundamental element of the hidden layer. The sequence of input in the time frame given as, {x (1), x (2), ⋯⋯⋯, x (M)} $\in R^{KxM}$, here x (τ) $\in R^{K}$ represents the feature vector for time step τ.

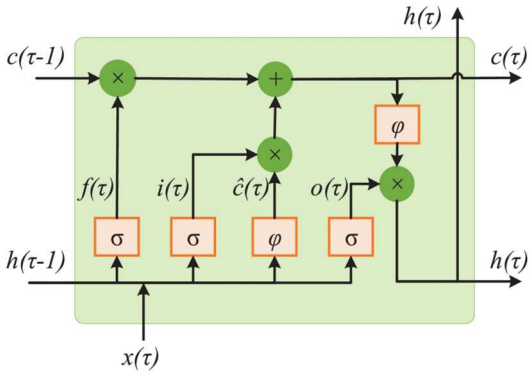

Fig. 4. Structural aspect of an LSTM neural network memory block.

The following equations (Eq. (1 – 3)) represents the gates of LSTM,

$$i_\tau = \sigma\,(w_i\,[h_{\tau-1}, x_t] + b_i) \quad (1)$$

$$f_\tau = \sigma\,(w_f\,[h_{\tau-1}, x_t] + b_f) \quad (2)$$

$$o_\tau = \sigma\,(w_o\,[h_{\tau-1}, x_t] + b_o) \quad (3)$$

where i_τ denotes the input gate, f_τ denotes the forget gate and o_τ denotes the output gate. w_i represents the weight in case of input gate, w_f represents weight in case of forget gate and w_o represents weight in case of output gate, each for 'x' number of neurons. Whereas σ is known as the sigmoid function. The output from the previous block of LSTM is represented by $h_{\tau-1}$. For the timestamp $\tau-1$, the input for the ongoing timestamp is x_t. The bias for the input, forget, and output gates is b_i b_f and b_o, respectively.

The input gate stores new instructions in the cell and further, for timeframe 'τ', the output gate provides activation to the LSTM block's final output, whereas the forget gate lets us know which information is not necessary and throw away.

$$\hat{c}_\tau = tanh\,(w_c\,[h_{\tau-1}, x_\tau] + b_c) \quad (4)$$

$$c_\tau = f_\tau * c_{\tau-1} + i_\tau * \hat{c}_\tau \quad (5)$$

$$h_\tau = o_t * tanh\,(c^\tau) \quad (6)$$

The cell state (memory) is represented by c_τ for timestamp (τ), $\hat{c}_\tau$ represents a candidate for cell state at timestamp (t). LSTM model is quite good for air quality prediction as the input data is sequence data with long-range dependencies.

C. Prediction of Particulate Matter

Then, at last, the cell state is filtered and passed to the activation function, which anticipates the portion for the output of the present LSTM unit at timestamp t. Then, h_{τ} output is passed from the present block of LSTM via the softmax layer to obtain the predicted output (y {τ}) from the present block. The multivariate LSTM model is implemented, and prediction is made for $PM_{2.5}$ with meteorological variables as secondary inputs. The model is trained with a validation split of 0.2 and a training split of 0.8. The forecasting horizon is dynamically changed from 1 to 30 to compare the forecast.

Statistical errors like The Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) are calculated to stabilise the results and check the model's accuracy. The square root of the average error is defined as RMSE (Eq. (7), while MAPE (Eq. (8) is the total of the individual absolute error divided by the demand (each period separately), which are given as,

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}} \quad (7)$$

$$MAPE = \frac{\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|}{n} \quad (8)$$

where, At represents actual or measured value, *Ft* represents forecasted value, and n indicates evaluated data point's number.

IV. Results and Discussions

This study uses the LSTM model to predict $PM_{2.5}$ for Singapore's north and south sites for the year 2019. The multivariate model is implemented with layers (1 layer) and neurons (50 neurons). The forecasting horizon is varied from 1 to 30 to observe the results for months across two sites. This means that the analysis is done using 1 day ahead, 3 days ahead, 5 days ahead, 7 days ahead, 14 days ahead, 21 days ahead and 30 days ahead for the two sites. Predictions for $PM_{2.5}$ are made with the assistance of input weather parameters.

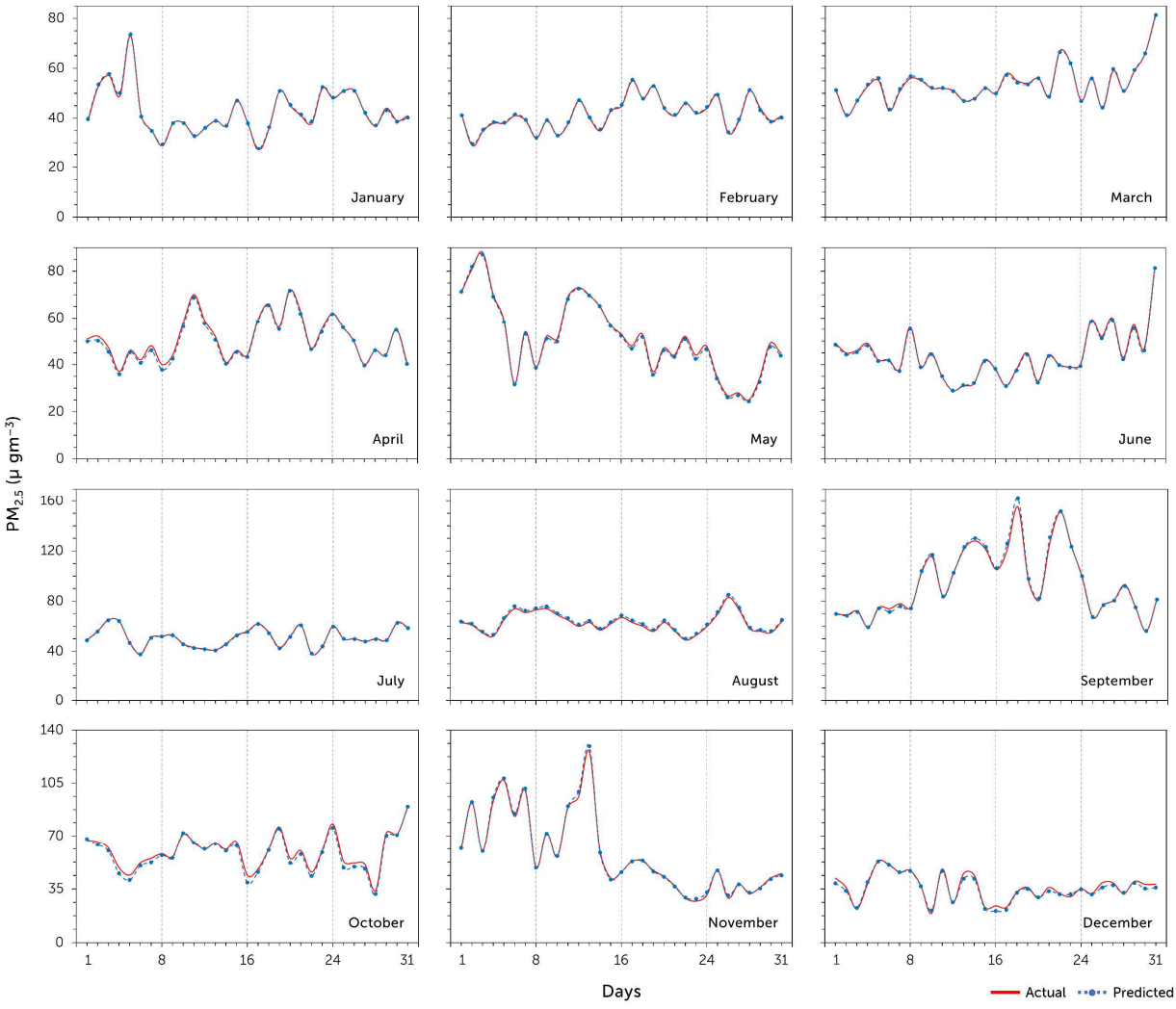

Fig. 5. Monthly predicted values of $PM_{2.5}$ for the north site in Singapore with 1 horizon. The red line represents actual values and the blue line represents the predicted values.

The predicted value of $PM_{2.5}$ is found to be varied from 20 to 160 µg m^{-3} per day for the north site, while 25 to 180 µg m^{-3} per day for the south site. The lowest and highest predicted values are observed for both sites during December and September, respectively. Table IV highlights the statistical errors for the predicted values of $PM_{2.5}$ for each site of horizon 3. Mostly, the statistical errors are significant for the monsoon season (JJA) for the north site and the winter season for the south site (DJF). The errors are relatively proportional to the number of horizons where the minimum RMSE is observed for horizon 5 (north site) and horizon 3 (south site). Likely, for MAPE minimum is horizon 3 (both the sites).

For 1 day ahead, RMSE is varied from 0.38 to 1.80 for the north site, relatively from 0.20 to 2.14 µg m^{-3} per day for the south site. Synoptically, the overall RMSE is less for the north site when compared to the south for horizon 1. Similar to RMSE, MAPE is varied from 0.80 to 2.77 µg m^{-3} per day for the north site, whereas from 0.27 to 1.96 µg m^{-3} per day for the south. The prediction for $PM_{2.5}$ with 1 day ahead over the selected sites of Singapore is shown in Fig. 5. For 1 day ahead, it can be further inferred that the north site exhibits the highest MAPE during December. On the other hand, the best forecasting is made during February (south site) with minimum RMSE and MAPE. The statistical errors for the horizon revealed the highest RMSE for September across the sites with 3.12 and 4.48 µg m^{-3} per day for north and south, respectively. MAPE is highest in winter with 8.01 (December, north) and 5.82 (November, south) µg m^{-3} per day. Matching with 1 day ahead, the pre-monsoon season (February) exhibited minimum statistical errors for 3 days ahead and are found to be in good fit with the observation compared to 1 day ahead predictions. The LSTM prediction for monthly $PM_{2.5}$ with 3 day ahead over the north site of Singapore is represented in Fig. 6.

Following the track of 3 days ahead, RMSE for 5 days ahead is observed with a maximum for November, with 3.767 and 3.93 µgm^{-3} per day for the north and south sites, respectively. Similarly, the north (south) site exhibited higher MAPE in December (November) with 7.83 (6.82) µg m^{-3} per day. Deviating from the earlier horizon's trend, the best prediction is made for the monsoon season (July) for the north site. Comparing the track of 1, 3 and 5, 3 days ahead is found to produce the best forecast with minor errors. For 7 days ahead, the RMSE is found to be maximum for the post-monsoon with values 5.89 (November, north) and 4.87 (October, south) µg m^{-3} per day and is dilutional towards the monsoon/winter season. But MAPE retains its highest errors for the winter season with 8.65 and 7.40 µg m^{-3} per day for north and south sites, respectively.

Whereas, the best-fitted forecast is observed for the monsoon season with minimal RMSE and MAPE. But, 7 days ahead is not the best-projected value when compared to 3 days ahead. The monthly statistical values for the sites for 1 and 3 days ahead are represented in Table IV. For 14 days ahead, the maximum RMSE is observed over the north (November) at 6.69 and south (September) at 13.34 µg m^{-3} per day. Similar to RMSE, MAPE exhibited lower significance with 10.09 and 11.14 µg m^{-3} per day for north (December) and south

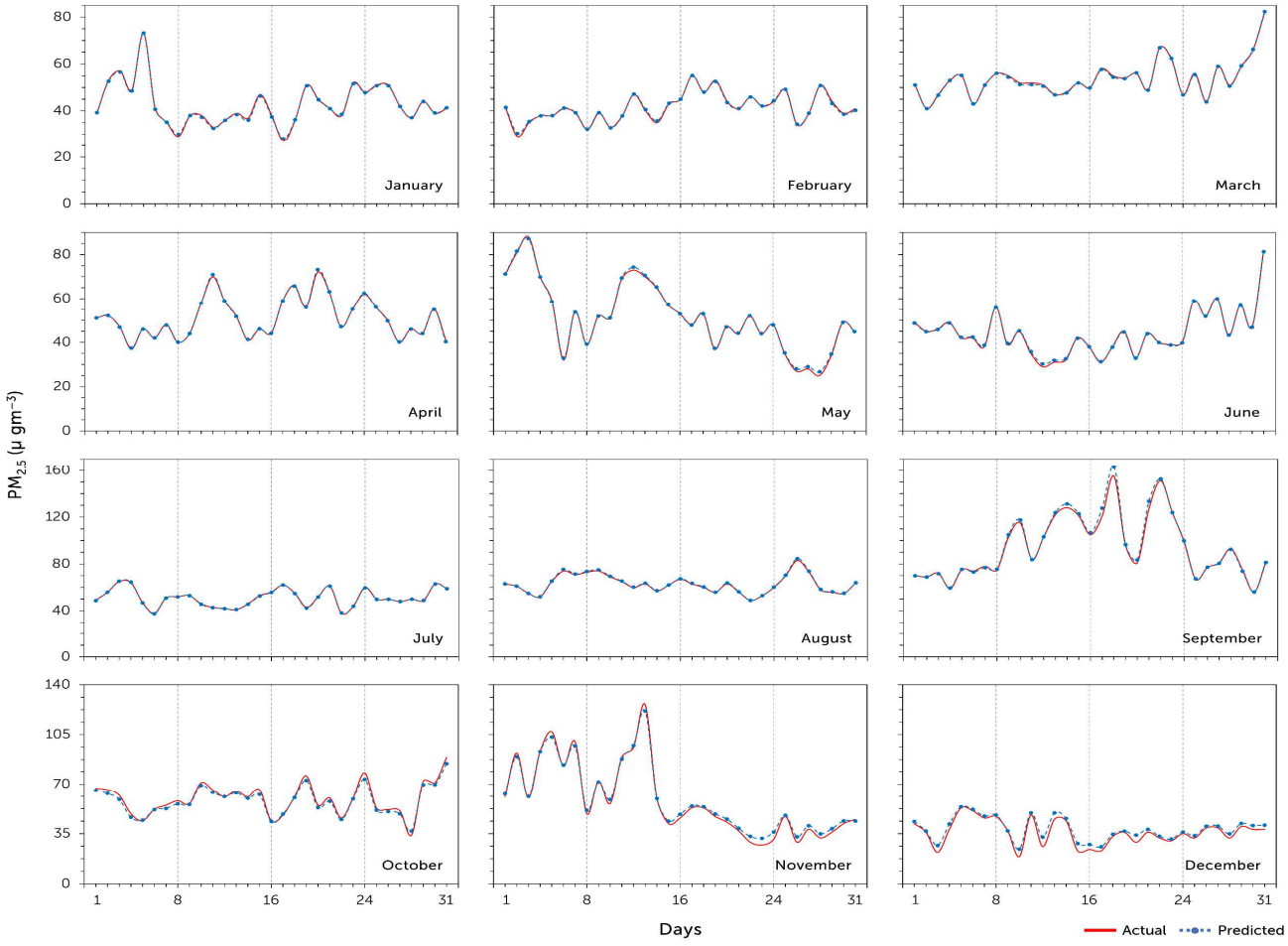

Fig. 6. Monthly predicted values of $PM_{2.5}$ for the north site in Singapore with 3 horizon. The red line represents actual values and the blue line represents the precdicted values.

(November), respectively. Though horizon 14 is significant during monsoon season, the errors offset the overall projections. The number of horizons is directly proportional to the insignificance of the projection, and this, in turn, is coherent for 21 days ahead as well. Higher error values of RMSE (MAPE) were observed over the winter season, with 11.51 (32.85) for north and 14.15 (25.89) μg m^{-3} per day for south sites of Singapore. Similarly, for 30 days ahead, the maximum error is observed for the post-monsoon season (September) with the variability of 15.48 (17.94) μg m^{-3} per day for north (south).

TABLE IV. STATISTICAL ERRORS FOR CORRESPONDING PREDICTED VALUES OF $PM_{2.5}$ FOR THE SITES IN SINGAPORE WITH HORIZONS 1 AND 3. SIGNIFICANT STATISTICAL FIGURES ARE HIGHLIGHTED (R: RMSE AND M: MAPE)

Month	Horizon 1				Horizon 3			
	North		South		North		South	
	R	M	R	M	R	M	R	M
Jan	**0.42**	**0.80**	**1.01**	**1.96**	0.48	0.96	0.53	0.93
Feb	0.53	1.21	**0.20**	**0.27**	0.30	0.54	0.15	**0.21**
Mar	1.03	1.63	0.43	0.59	0.50	0.81	0.33	0.36
Apr	0.72	1.23	0.51	0.60	0.37	**0.41**	0.61	0.82
May	0.78	1.59	0.79	1.85	0.66	1.17	0.94	2.13
Jun	0.38	0.83	0.34	0.75	0.39	0.81	**0.37**	0.77
Jul	1.12	1.84	0.38	0.59	**0.24**	0.36	0.38	0.52
Aug	1.53	2.31	0.65	0.54	0.53	0.58	0.54	0.52
Sep	1.80	1.22	2.14	1.22	**3.12**	1.67	**4.49**	3.25
Oct	1.43	2.12	0.80	1.16	2.31	3.24	2.38	3.03
Nov	0.64	1.32	0.67	1.25	2.40	4.86	2.76	**5.82**
Dec	**1.14**	**2.77**	0.76	1.70	2.97	**8.02**	1.41	3.38

From the analysis, synoptically, it can be inferred that 30 days ahead holds the maximum error variability, whereas 3 days ahead is considered a good fit in predictions. In general, the south site of Singapore exhibited a better forecast during the post-winter (February) in accordance with the observation, whereas higher statical errors are observed during the onset of winter (December).

As it is observed that increasing horizon is leading to higher errors and worst prediction, this is due to overfitting of the model with increment in batch size. LSTM network uses a reduced number of hidden units, lags, and also a minimum number of training iterations; when training iteration is increased, learning increases and hence better forecast is obtained. But at a certain point, LSTM begins to overlearn; hence forecast accuracy maximizes at that number only. Likewise, the results are obtained for horizon 3, where accuracy is maximum. Beyond horizon 3, the estimating capabilities get weak, and hence model gets stuck to the input responses from input data more often and leads to the increment of errors. And this is observed for the horizons with 21 and 30, where estimating capabilities are weak, and MAPE and RMSE are found to be maximum.

For horizon 14, the maximum RMSE is observed over the north (November) at 6.69 and south (September) at 13.34 μg m^{-3} per day. Similar to RMSE, MAPE exhibits lower significance with 10.09 and 11.14 μg m^{-3} per day for north (December) and south (November), respectively. Though 14 days ahead is significant during monsoon season, the errors offset the overall projections. The number of horizons is directly proportional to the insignificance of the projection, and this, in turn, is coherent for horizon 21 as well.

Higher error values of RMSE (MAPE) are observed over the winter season with 11.51 (32.85) for north and 14.15 (25.89) μg m^{-3} per day for south sites of Singapore. Similarly, for 30 days ahead, the maximum error is observed for the post-monsoon season (September) with the variability of 15.48 (17.94) μg m^{-3} per day for north (south). So, forecast accuracy is mainly dependent on the number of hidden units, training iterations, and value of lag; depending on each other, the optimum values of these three increase the accuracy of the model.

V. CONCLUSION

This work has presented an LSTM-based model to forecast daily $PM_{2.5}$ levels over Singapore over multiple time horizons, a major pollutant to the atmosphere, over the selected sites of Singapore. Firstly, the correlation between the input variables is obtained and then the model is trained with 50 neurons and 1 layer. Multiple horizon forecasts are obtained with changing the horizons from 1 to 30 concerning the model's accuracy and forecasting capabilities. The statistical errors RMSE and MAPE were computed to evaluate the model. For the sites of Singapore, it has been found that better results are obtained for the 3-day ahead forecast, whereas the model is non-significant for the maximum horizons like 21 and 30. When the number of horizon increases, the model starts overfitting, which reduces the forecast's accuracy.

From the forecasted analysis, it is inferred that $PM_{2.5}$ is better projected in the pre-monsoon season (mostly in February), and a comparatively weaker forecast is observed for the winter season. On comparing our model performance with the existing literature [18], the proposed algorithm can perform predictions with marginal bias, exhibiting the suitability to forecast air quality in real time. This preliminary study in the projection of $PM_{2.5}$ over Singapore is anticipated to assist policymakers in framing better decisions for preventing and timely air pollution control over Singapore.

ACKNOWLEDGMENT

The authors would like to acknowledge the National Institute of Technology Rourkela because of handling the lab facilities in conducting this research.

DATA AVAILABILITY STATEMENT

The data used for the analysis and the algorithm can be found in the online data repository: https://doi.org/10.5281/zenodo.8079096 [23].

REFERENCES

[1] Health Care Equality Index, 2019, https://assets2.hrc.org/files/assets/resources/HEI-2019-FinalReport.pdf.

[2] Health World Health Organisation, 2016, https://www.who.int/gho/publications/world_health_statistics/2016/EN_WHS2016_TOC.pdf.

[3] Bedi, T.K. and Bhattacharya, S.P., 2021. An Investigative Study on Perceived Indoor Air Quality During COVID-19 Lockdown in India. Journal of The Institution of Engineers (India): Series A, 102(4), pp.885-900.

[4] Raju, K.V., Ravindra, A., Manasi, S., Smitha, K.C. and Srinivas, R., 2018. Urban Environmental Governance in India.

[5] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," *Frontiers in Public Health*. 2020, doi: 10.3389/fpubh.2020.00014.

[6] M. Gupta *et al.*, "AI-enabled COVID-19 outbreak analysis and prediction: Indian states vs. union territories," *Comput. Mater. Contin.*, 2021, doi: 10.32604/cmc.2021.014221.

[7] Zhang, Xuan, Xun Liang, Aakas Zhiyuli, Shusen Zhang, Rui Xu, and Bo Wu. "At-lstm: An attention-based lstm model for financial time series prediction." In IOP Conference Series: Materials Science and Engineering, vol. 569, no. 5, p. 052037. IOP Publishing, 2019.

[8] Navares, Ricardo, and José L. Aznarte. "Predicting air quality with deep learning LSTM: Towards comprehensive models." Ecological

Informatics 55 (2020): 101019.

[9] A. Hussain *et al.*, "Waste management and prediction of air pollutants using IoT and machine learning approach," *Energies*, 2020, doi: 10.3390/en13153930.

[10] Abdel-Nasser, M. and Mahmoud, K., 2019. Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Computing and Applications*, *31*(7), pp.2727-2740.

[11] Prasanth, S., Singh, U., Kumar, A., Tikkiwal, V.A. and Chong, P.H., 2021. Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach. Chaos, Solitons & Fractals, 142, p.110336.

[12] Chandola, D., Gupta, H., Tikkiwal, V.A. and Bohra, MK, 2020. Multi-step ahead forecasting of global solar radiation for arid zones using deep learning. Procedia Computer Science, 167, pp.626-635.

[13] Ding, G. and Qin, L., 2020. Study on the prediction of stock price based on the associated network model of LSTM. International Journal of Machine Learning and Cybernetics, 11(6), pp.1307-1317.

[14] Livieris, I.E., Pintelas, E. and Pintelas, P., 2020. A CNN–LSTM model for gold price time-series forecasting. Neural computing and applications, 32(23), pp.17351-17360.

[15] Xiao, Y., Yin, H., Zhang, Y., Qi, H., Zhang, Y. and Liu, Z., 2021. A dual‐stage attention‐based Conv‐LSTM network for spatio‐temporal correlation and multivariate time series prediction. International Journal of Intelligent Systems, 36(5), pp.2036-2057.

[16] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2923707.

[17] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents," *Comput. Oper. Res.*, 2019, doi: 10.1016/j.cor.2018.02.021.

[18] Y. T. Tsai, Y. R. Zeng, and Y. S. Chang, "Air pollution forecasting using rnn with lstm," 2018, doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00178.

[19] R. Casado-Vara, A. M. del Rey, D. Pérez-Palau, L. De-La-fuente-valentín, and J. M. Corchado, "Article web traffic time series forecasting using LSTM neural networks with distributed asynchronous training," *Mathematics*, 2021, doi: 10.3390/math9040421.

[20] M. L. Shen, C. F. Lee, H. H. Liu, P. Y. Chang, and C. H. Yang, "Effective multinational trade forecasting using LSTM recurrent neural network," *Expert Syst. Appl.*, 2021, doi: 10.1016/j.eswa.2021.115199.

[21] Hansen, A.B., Witham, C.S., Chong, W.M., Kendall, E., Chew, B.N., Gan, C., Hort, M.C. and Lee, S.Y., 2019. Haze in Singapore–source attribution of biomass burning PM 10 from Southeast Asia. Atmospheric Chemistry and Physics, 19(8), pp.5363-5385.

[22] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, 2018, doi: 10.1016/j.energy.2018.01.177.

[23] A. Tripathi, M. Hari, V. A. Tikkiwal, B. Tyagi and A. Kumar, "Multi-horizon forecasting of average daily PM2.5 concentration using LSTM: A case study of air pollution in Singapore," [Data set]. *Zenodo*., 2023, doi: 10.5281/zenodo.8079237.

[24] X. Yan, Z. Zang, N. Luo, Y. Jiang, and Z. Li, "New interpretable deep learning model to monitor real-time PM2.5 concentrations from satellite data," *Environ. Int.*, 2020, doi: 10.1016/j.envint.2020.106060.

[25] S. V. Belavadi, S. Rajagopal, R. Ranjani, and R. Mohan, "Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks," 2020, doi: 10.1016/j.procs.2020.03.036.

[26] M. S. Hossain and H. Mahmood, "Short-term photovoltaic power forecasting using an LSTM neural network and synthetic weather forecast," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3024901.