Journal of Cleaner Production 209 (2019) 134–145

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts

Yanlai Zhou [a], Fi-John Chang [a, *], Li-Chiu Chang [b], I-Feng Kao [a], Yi-Shin Wang [a]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, 10617, Taiwan, ROC
[b] Department of Water Resources and Environmental Engineering, Tamkang University, New Taipei City, 25137, Taiwan, ROC

ARTICLE INFO

Article history:
Received 2 July 2018
Received in revised form
26 September 2018
Accepted 22 October 2018
Available online 24 October 2018

Keywords:
Multi-output LSTM
Deep learning
Artificial intelligence (AI)
Multi-step-ahead forecast
Air quality
Taipei city

ABSTRACT

Timely regional air quality forecasting in a city is crucial and beneficial for supporting environmental management decisions as well as averting serious accidents caused by air pollution. Artificial Intelligence-based models have been widely used in air quality forecasting. The Shallow Multi-output Long Short-Term Memory (SM-LSTM) model is suitable for regional multi-step-ahead air quality forecasting, while it commonly encounters spatio-temporal instabilities and time-lag effects. To overcome these bottlenecks and overfitting issues, this study proposed a Deep Multi-output LSTM (DM-LSTM) neural network model that were incorporated with three deep learning algorithms (i.e., mini-batch gradient descent, dropout neuron and L2 regularization) to configure the model for extracting the key factors of complex spatio-temporal relations as well as reducing error accumulation and propagation in multi-step-ahead air quality forecasting. The proposed DM-LSTM model was evaluated by three time series of $PM_{2.5}$, $PM_{10,}$ and NO_x simultaneously at five air quality monitoring stations in Taipei City of Taiwan. Results indicated that the loss function values (mean-square-error) of the SM-LSTM and DM-LSTM models in the testing stages at horizon t+4 were 0.87 and 0.72, respectively. The G_{bench} values of the DM-LSTM model in the testing stages for $PM_{2.5}$, $PM_{10,}$ and NO_x reached 0.95 at horizon t+1 and exceeded 0.81 at horizon t+4, respectively. Results demonstrated that the proposed DM-LSTM model incorporated with three deep learning algorithms could significantly improve the spatio-temporal stability and accuracy of regional multi-step-ahead air quality forecasts.

1. Introduction

Exposure to ambient air pollution is a primary environmental risk factor in relation to adverse health impacts (Apte et al., 2015; Liu et al., 2017). Fine particulate matter ($PM_{2.5}$ and PM_{10}, i.e., particles smaller than 2.5 or 10 μm) and nitrogen oxide (NO_x) are the dominant components of ambient air pollution associated with booming urban development (Li et al., 2017, 2018; Lin and Zhu, 2018). To date, epidemiological investigations and studies demonstrated that some air pollution-related diseases were associated with exposure to $PM_{2.5}$, $PM_{10,}$ and NO_x (Reggente et al., 2014; Wang et al., 2016; Wu et al., 2018a,b). In addition, these air pollutants were acknowledged as typical representatives of particle number concentration in urban air quality (Li et al., 2018a,b; Wu et al., 2018a,b). Real-time air quality information is of great importance to air pollution control and human health protection from air pollution (Ni et al., 2017). To support environmental management decisions and avert serious accidents caused by air pollution, air quality forecasting is becoming more and more essential not only to better govern the trend of air pollution variation but to provide timely and comprehensive environmental quality information (Pournazeri et al., 2014; Yang and Christakos, 2015; Corani and Scanagatta, 2016; Lauret et al., 2016; Wakeel et al., 2017; Van et al., 2018; Yang et al., 2018).

A wide variety of methods have been used to forecast or predict regional air quality. These studies primarily branched out into two major classes: physical-based and data-driven methods. Physical-based models like dispersion and chemical transport models have still been under development as a result of uncertainties in relation to source inventories and the chemical and dynamical mechanisms of aerosols in atmosphere (Afzali et al., 2017; Vijayaraghavan et al., 2016; Jiang et al., 2018; Karambelas et al., 2018; Pisoni et al., 2018). Data-driven models have leant upon the empirical or statistical relationship between air quality observations and other affecting

* Corresponding author.
E-mail address: changfj@ntu.edu.tw (F.-J. Chang).

https://doi.org/10.1016/j.jclepro.2018.10.243
0959-6526/

factors (Ausati and Amanollahi, 2016; Gong and Ordieres, 2016; Gao et al., 2018). Among data-driven models, artificial neural networks (ANNs), a crucial branch of Artificial Intelligence (AI), have been utilized frequently to predict environmental parameters, especially for water and air quality (e.g., Feng et al., 2015; Chang et al., 2015, 2016; Taghavifar et al., 2016; Taylan, 2017; Nieto et al., 2017). For instance, the backpropagation neural networks (BPNN), the radial basis function (RBF), the Elman recurrent neural network, the nonlinear autoregressive with exogenous inputs neural network (NARX), the adaptive-network-based fuzzy inference system (ANFIS) and the support vector machine (SVM) have been widely applied to modelling air quality (Voukantsis et al., 2011; Reisen et al., 2014; Prasad et al., 2016; Yeganeh et al., 2018; Nieto et al., 2018). Nevertheless, these models usually encounter a common drawback of under-predicting particulate matter concentrations under the conditions of very high concentrations that would impose the most adverse effects on human health. Therefore, more prior knowledge and sophisticated modelling techniques are needed to capture the abrupt changes in particulate matter concentrations. It was noticed that the abovementioned methods were usually adopted to construct site-specific data-driven models for individual air quality monitoring station, in disregard of the potential nonlinear spatial correlation among air quality monitoring stations. Bearing this in mind as a motivation, multi-output data-driven models adopted in forecasting would generally be the instance that the underlying nonlinear correlation among output variables could be extracted to improve forecast accuracy (Nguyen et al., 2012; Li et al., 2016). The demand for multi-step-ahead and multi-output air quality forecasting increased modelling difficulty when traditional shallow neural network models were implemented. Recently, the Long Short-Term Memory (LSTM) neural network, serving as an advanced component of deep learning neural networks, has been applied with success to image classification, natural language processing, Internet of Things (IoT), machine translation, and prediction. (Krizhevsky et al., 2012; Ballesteros et al., 2017; Greff et al., 2017; Zhang et al., 2018a,b). Wei et al. (2017) used the convolution-LSTM-based deep neural network to analyze and predict spatiotemporal data and promote text transfer learning research. Hinton et al. (2006, 2012) utilized deep learning neural networks configured with network architectures of multiple hidden layers to capture the inherent features of data layer-by-layer without prior knowledge, which produced good performance in time series forecasting. Therefore, it is imperative to conduct in-depth research on the multi-output data-driven models constructed over deep learning neural networks for improving forecast reliability and accuracy through tackling the complexity and challenges encountered in regional multi-step-ahead air quality forecasting.

This study was explored with two primary foci: (1) developing a deep learning-based multi-output LSTM neural network (DM-LSTM) model to make regional multi-ahead-step forecasts at multiple outputs simultaneously; and (2) integrating three deep learning algorithms to train the DM-LSTM model for overcoming the bottlenecks of instability and overfitting. The proposed DM-LSTM model with h ($\geq$2) hidden layers was trained by a composition of three deep learning algorithms for extracting the complex spatio-temporal patterns among meteorological inputs, air quality inputs and multiple air quality outputs at different air quality monitoring stations. The reliability and accuracy of the proposed model were assessed by a study case of the regional multi-ahead-step air quality ($PM_{2.5}$, PM_{10}, and NO_x) forecasts in Taipei City of Taiwan.

2. Methodology

This paper proposed a deep learning-based multi-output LSTM neural network model (DM-LSTM) for improving the multi-step-ahead forecast accuracy of multiple outputs, where the model was trained by a composition of three deep learning algorithms with weight adjustment. Fig. 1 illustrated the architectures of the original LSTM unit (Fig. 1 (a)), the Shallow Multi-output LSTM neural network (SM-LSTM) model with one hidden layer (Fig. 1 (b)), and the proposed Deep Multi-output LSTM neural network (DM-LSTM) model with h ($\geq$2) hidden layers (Fig. 1 (c)). The SM-LSTM model served as a benchmark in this study. The methods used in this study were briefly introduced as follows.

2.1. Long Short-Term Memory (LSTM) model

As one of the popular recurrent neural networks, the LSTM neural network with the internal self-looped cell was first proposed by Hochreiter and Schmidhuber (1997), which promoted the ability to memorize the long (static) term and short (recurrent) term dynamic characteristics of time series (Hochreiter, 1998). The description of the LSTM unit was given in Appendix A.

The SM-LSTM model was introduced as follows.

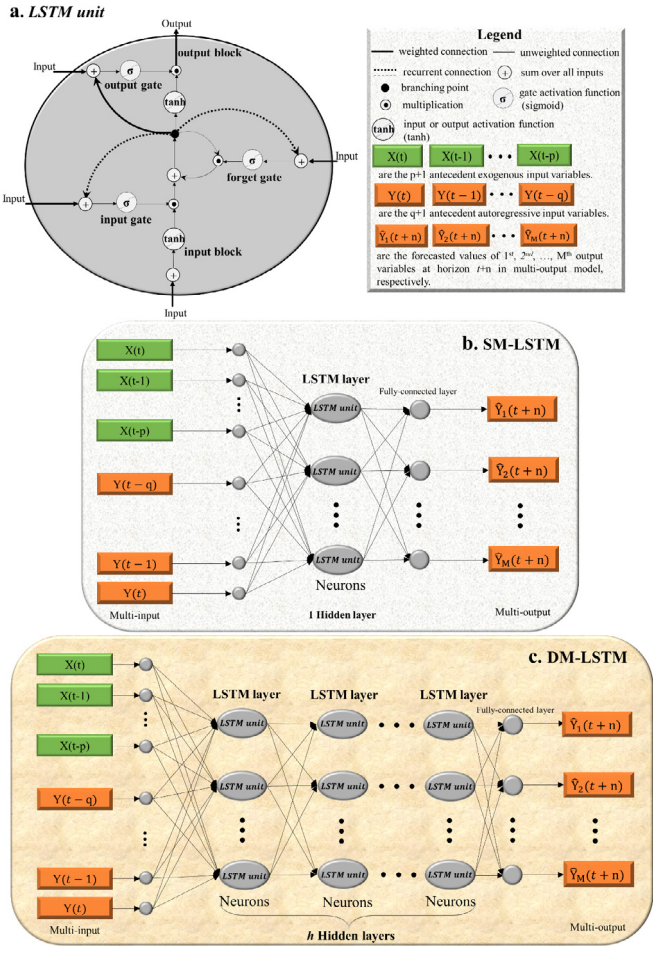

Fig. 1. Architectures of Multi-output Long Short-Term Memory neural network models (M-LSTM). a. LSTM unit. b. Shallow learning-based M-LSTM model with one hidden layer (SM-LSTM). c. Deep learning-based M-LSTM model with h hidden layers (DM-LSTM).

$$\widehat{\mathrm{Y}}(t+\mathrm{n}) = f\left((\mathrm{X}_t, \mathrm{Y}_t), \mathrm{H}^1\right) = T\left((\mathrm{X}_t, \mathrm{Y}_t)\middle|\mathrm{H}^1\right) \cdot T\left(\mathrm{H}^1\right) \quad (1)$$

where X_t and Y_t are the observed exogenous and autoregressive model input variables, respectively. $\widehat{\mathrm{Y}}(t+\mathrm{n}) = [\widehat{\mathrm{Y}}_1(t+\mathrm{n}), \widehat{\mathrm{Y}}_2(t+\mathrm{n}), \ldots, \widehat{\mathrm{Y}}_{\mathrm{M}}(t+\mathrm{n})]$ are the 1st, 2nd, …, M^{th} multi-output variables at horizon $t+\mathrm{n}$. H^1 denotes the only hidden layer in the SM-LSTM model. $f((\mathrm{X}_t, \mathrm{Y}_t), \mathrm{H}^1)$ is the mapping function between model input and output. $T((\mathrm{X}_t, \mathrm{Y}_t)|\mathrm{H}^1)$ is the conditional probability function between model input and the hidden layer. $T(\mathrm{H}^1)$ is the transform function of the hidden layer.

The DM-LSTM model was described as follows.

$$\begin{aligned}\widehat{\mathrm{Y}}(\mathrm{t}+\mathrm{n}) &= f\left((\mathrm{X}_t, \mathrm{Y}_t), \mathrm{H}^1, \mathrm{H}^2, \ldots, \mathrm{H}^h\right) \\ &= T\left((\mathrm{X}_t, \mathrm{Y}_t)\middle|\mathrm{H}^1\right) \cdot T\left(\mathrm{H}^1\middle|\mathrm{H}^2\right) \cdot \ldots, \cdot T\left(\mathrm{H}^{h-1}\middle|\mathrm{H}^h\right) \cdot T\left(\mathrm{H}^h\right)\end{aligned} \quad (2)$$

where H^1, H^2, …, H^h are the 1st, 2nd, …, hth hidden layers in the DM-LSTM model. Under the precondition with the same number of input variables (N_{in}), the number of neurons (m) in every hidden layer and output variables (N_{out}), the total number of parameters in the SM-LSTM model with one hidden layer equals to the sum of $(\mathrm{N}_{in} \times m + m \times \mathrm{N}_{out})$ weight parameters and $(m + \mathrm{N}_{out})$ bias parameters, while the total number of parameters in the DM-LSTM model with h hidden layers equals to the sum of $(\mathrm{N}_{in} \times m + (h-1) \times m^2 + m \times \mathrm{N}_{out})$ weight parameters and $(h \times m + \mathrm{N}_{out})$ bias parameters, respectively. In other words, the DM-LSTM model would have extra $((h-1) \times m^2 + (h-1) \times m)$ parameters owing to h hidden layers.

The comparison between SM-LSTM and DM-LSTM models constructed in this study was summarized as: (1) to extract the inherent features of data from the lowest to the highest levels layer-by-layer, the former used a shallow neural network (SNN) with one hidden layer while the latter used a deep neural network (DNN) with h ($\geq$2) hidden layers; and (2) on account of h hidden layers, the latter had more parameters (i.e.$(h-1) \times m^2 + (h-1) \times m$) than the former did. In addition, the stochastic gradient descent algorithm (SGD) was commonly used to optimize the parameters of a multi-output LSTM model with one hidden layer (Nakama, 2009). The SGD applied to DM-LSTM models with more than one hidden layer usually encountered bottlenecks of instability and overfitting (Hinton et al., 2012). In other words, DM-LSTM models would demand for more auxiliary deep learning techniques to increase model stability and mitigate overfitting.

2.2. Deep learning algorithms

DNNs are considered suitable for modelling the non-linear spatio-temporal pattern upon time series, while it is easy for them to trigger overfitting problems if the number of network parameters is large (Hinton et al., 2012). To mitigate overfitting and increase stability, three deep learning algorithms, i.e., mini-batch gradient decent (MBGD) algorithm, dropout neuron algorithm, and L2 regularization algorithm, were integrated to train the DM-LSTM model in this study. To be more precise, the advantages of the proposed method lay in two folds: the use of the MBGD algorithm aimed at mitigating model instability while the use of the dropout neuron and L2 regularization algorithms aimed at the avoidance of overfitting. The three deep learning algorithms were briefly described as below.

2.2.1. MBGD algorithm

The full batch gradient descent algorithm using all training datasets in each iteration and the SGD algorithm using one training dataset in each iteration are common practice for training SNN models (Rumelhart et al., 1985). The former benefits from better convergence but suffers from slower computation speed due to the need to observe the whole training datasets in every iteration. The latter benefits from faster computation speed but suffers from inferior convergence. To overcome the drawbacks of the SGD and the full batch gradient decent algorithms, the MBGD algorithm takes the advantages of both algorithms and performs an update on parameters for every mini-batch of training datasets (Nakama, 2009), which reduces the variance of the updates on parameters and would usually produce more stable convergence (Qian et al., 2015). Therefore, the MBGD algorithm was applied to optimizing the parameters of the DM-LSTM model in this study.

2.2.2. Dropout neuron algorithm

The dropout algorithm randomly discards some neurons with probability p in the hidden layer when training a neural network for preventing the co-adaptation of neurons (Baldi and Sadowski, 2014). From the perspective of the reduction in model structure complexity, the dropout neuron algorithm is also considered as an effective method to handle the overfitting problems of DNN models (Hinton et al., 2012; Srivastava et al., 2014). Through dropping out neurons, the parameters of the DM-LSTM model proposed in this study were updated by a backpropagation algorithm such that the connections of the survived neurons became more stable. In other words, only the survived neurons were trained at every iteration. Hence, the dropout neuron algorithm could transform a fully-connected hidden layer into a partially-connected one for preventing the DNN model from depending overmuch on deterministic neurons in the hidden layers and consequently could mitigate the co-adaptability of neurons (Zhang et al., 2018a,b).

2.2.3. L2 regularization algorithm

The L2 regularization algorithm is usually adopted to optimize the weight parameters of data-driven models for avoiding overfitting (Chang et al., 2010; Kabán, 2013; Bilgic et al., 2014; Nielsen, 2015; Wang and Cao, 2017). As a penalty, the L2 regularization algorithm adds the sum of the absolute values of weight parameters to the loss function (or the objective function) through gradient descent calculation. The mean-square-error (MSE) is commonly used as the loss function (L_0) in the gradient descent calculation. The MSE was described as follows.

$$\mathrm{L}_0 = \mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{t=1}^{\mathrm{N}} \left(\mathrm{Y}(t) - \widehat{\mathrm{Y}}(t)\right)^2 \quad (3)$$

where $\mathrm{Y}(t)$ and $\widehat{\mathrm{Y}}(t)$ are the matrixes of observed and forecasted multi-output variables at the tth time, respectively. N is the number of time steps. From the perspective of updated weight parameters, the MSE with L2 regularization (Fig. 2 (c)) was used as the loss function (L) in the gradient descent calculation in this study, described as follows.

$$\mathrm{L} = \mathrm{L}_0 + \frac{\lambda}{2\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \omega_i^2 = \mathrm{MSE} + \frac{\lambda}{2\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \omega_i^2 \quad (4)$$

where L is the loss function with L2 regularization. $\frac{\lambda}{2\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \omega_i^2$ is the L2 regularization for weight parameters. λ is the coefficient of L2 regularization, and $\lambda > 0$. ω is the matrix of weight parameters.

The partial derivative of the loss function L to ω was described

Fig. 2. Flow diagram of the DM-LSTM trained by a composition of: a. Mini-Batch Gradient Decent algorithm (MBGD); b. Dropout neuron algorithm (LSTM units colored in white are dropped out randomly); and c. L2 regularization algorithm.

below

$$\frac{\partial L}{\partial \omega} = \frac{\partial L_0}{\partial \omega} + \frac{\lambda}{N}\omega \tag{5}$$

Thus, in each gradient computation, the weight parameter matrix ω could be updated by the following equations.

$$w \rightarrow w - \alpha\left(\frac{\partial L_0}{\partial w} + \frac{\lambda}{N}w\right) \tag{6a}$$

$$w - \alpha\left(\frac{\partial L_0}{\partial w} + \frac{\lambda}{N}w\right) = \left(1 - \alpha\frac{\lambda}{N}\right)w - \alpha\frac{\partial L_0}{\partial w} \tag{6b}$$

where α is the learning rate in gradient descent calculation, and $\alpha > 0$. Accounting for the positive parameters α, λ and N, the values in the weight parameter matrix ω tend to decrease in the iteration process.

In brief, the MBGD algorithm has the capability for overcoming the instability problem of DNN models, while the dropout neuron and L2 regularization algorithms have the ability for coping with the overfitting problems of DNN models at the same time. The following section described how to integrate the three deep learning algorithms for training the DM-LSTM model.

2.3. Training process of the LSTM model

In this study, the training process of the DM-LSTM model was executed with the MBGD, the dropout neuron, and the L2 regularization algorithms. Fig. 2 showed the flow diagram of the DM-LSTM model training process. The implementation procedure was described as follows.

Step 1: Divide datasets into training and testing datasets. Initialize the parameters of the LSTM. Set the maximal epoch, the number of hidden layers and the number of neurons.

Step 2: Identify a mini-batch of data based on training datasets through the trial-and-error procedure (Fig. 2(a)). The mini-batch size usually ranges between 32 (= 2^5) and 1024 (=2^{10}), which needs to be adjusted application by application for making sure the mini-batch size is suitable for the Central Processing Unit (CPU) or Graphic Processing Unit (GPU) memory (Nakama, 2009). Hence, the values of the batch size in this study were set as 2^5 (minimal value), 2^6, 2^7 and 2^8 (maximal value), respectively.

Step 3: Implement the routines of the dropout neuron and L2 regularization algorithms:

(3a) Check whether an epoch is completed. If an epoch is not completed, the number of neurons in the hidden layer is adjusted with the dropout probability p (Fig. 2 (b)).

(3b) Compute the loos function in terms of MSE with the L2 regularization algorithm (Eq. (10), Fig. 2(c)). Then, execute the gradient descent calculation, update weights and bias, and utilize the next mini-batch to repeat Step 3. If an epoch is completed, proceed to the next step.

Step 4: Terminate the computation process subject to the stopping criteria (early stopping or the maximal epoch E_{max}). If the value of the loss function does not decrease over 100 consecutive epochs, forecast accuracy can no longer be enhanced, which triggers the computation to stop. If the maximal number of epochs is reached, the training process stops. Otherwise, update the epoch, and repeat Steps 2 and 3.

Output: Save the optimized parameters of the DM-LSTM model, including the maximal epoch (E_{max}), the number of neurons, the learning rate (α), the mini-batch size, the dropout probability (p), the coefficient (λ) of L2 regularization, the weight vector and the bias vector.

3. Study area and materials

With the fast-growing economy and population, air quality deterioration in Taiwan has become highly problematic in recent years. Taipei City with an area of 272 km^2 serves as the center of politics, commerce, and culture in Taiwan. The population of the city reached 2.69 million in 2016. People across Taipei City nowadays undergo a great possibility of exposure to high-level invasion of air pollutants (e.g., $PM_{2.5}$, PM_{10}, and NO_x). Therefore, healthy and green urban development demands for accurate multi-step-ahead forecasts of $PM_{2.5}$, PM_{10} and NO_x concentrations such that regional air quality can be handled and controlled adequately.

Fig. 3 illustrated the locations of Taipei City, five air quality monitoring stations and sixteen meteorological monitoring stations in the study area. Regarding the air quality monitoring stations, Stations A1 (Yong-He) and A2 (San-Chong) are traffic stations (i.e., stations located in areas of heavy traffic), Stations A3 (Song-Shan) and A4 (Shi-Lin) are general stations, and Station A5 (Yang-Ming) is a park station (i.e., a station located in a park). This study employed hourly data of eight air quality factors ($PM_{2.5}$, PM_{10}, O_3, NO_x, NO_2, NO, SO_2, CO) and five meteorological factors (rainfall, temperature, wind speed, wind direction, and relative humidity) collected from 2010 to 2016 (7 years) in the study area. In this study, air quality data were extracted from the Environmental Protection Administration in Taiwan (https://taqm.epa.gov.tw/taqm/tw/default.aspx in Chinese), and meteorological data were extracted from the Central Weather Bureau in Taiwan (https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp in Chinese). A total of 61,368 (=[(2 × 366)+(5 × 365)] × 24) hourly datasets were used in this study, where 35,064 data (4 years) were used for model training while the remaining 26,304 data (3 years) were used for model

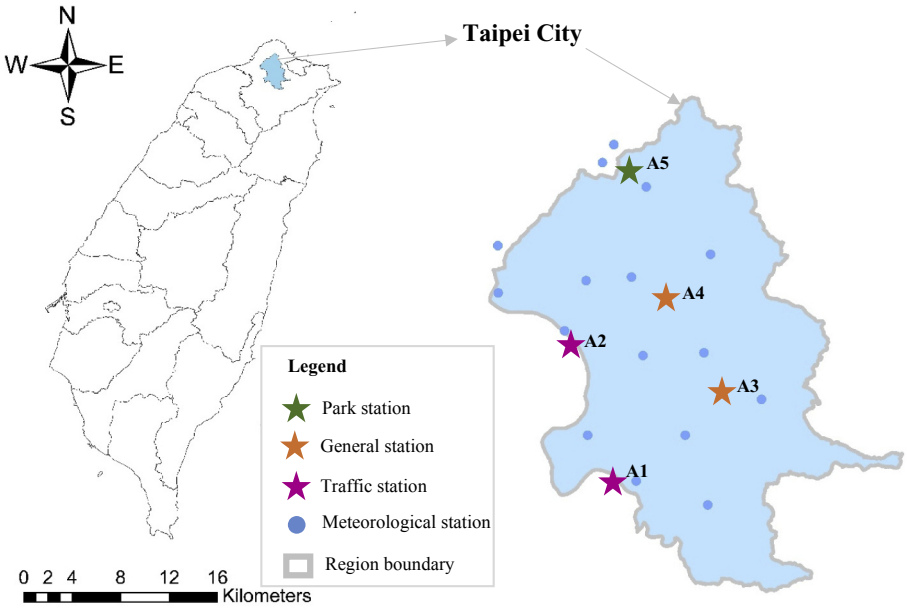

Fig. 3. Distribution of air quality and meteorological monitoring stations in Taipei City. Stations A1 (Yong-He) and A2 (San-Chong) are traffic stations (i.e. stations located in areas of heavy traffic). Stations A3 (Song-Shan) and A4 (Shi-Lin) are general stations. Station A5 (Yang-Ming) is a park station (i.e. a station located in a park).

testing. The Kendall tau coefficient (Maidment, 1993) was adopted to identify the highest correlation concerning time lags between input variables (meteorological and air quality factors) and output variables ($PM_{2.5}$, PM_{10}, and NO_x).

To reduce the negative effect of the different scales of input data on model's learning ability, all thirteen input variables were standardized to the same scale. For obtaining stable convergence in the developed model, the normal standardization was applied to data pre-processing. The standardization formula was defined as follows.

$$X^*(t) = \frac{X(t) - \overline{X}}{\sigma} \tag{7}$$

where $X^*(t)$ is the normal standardization for input data in the tth time. $\overline{X}$ and σ are the average and standard deviation of input data, respectively. The root-mean-square-error (RMSE) and the goodness-of-fit with respect to the benchmark (G_{bench}) were carried out for comparison. The RMSE and G_{bench} were defined as follows.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{Y}(t) - Y(t)\right)^2}, \quad RMSE \geq 0 \tag{8}$$

$$G_{bench} = \left(1 - \frac{\sum_{i=1}^{T}\left(\widehat{Y}(t) - Y(t)\right)^2}{\sum_{i=1}^{T}\left(Y(t) - Y_{bench}(t)\right)^2}\right) \times 100\%, \quad G_{bench} \leq 100\% \tag{9}$$

where $\widehat{Y}(t)$ and $Y(t)$ is the forecasted and observed values of the output variable at the tth time, respectively. $Y_{bench}(t)$ is the observed data shifted backwards by one or more time lags, e.g., for the nth-step-ahead forecast, $Y_{bench}(t) = Y(t-n)$.

Table 1 presented the statistic indexes of $PM_{2.5}$, PM_{10} and NO_x concentrations at five air quality monitoring stations in different seasons. It was noticed that the statistic indexes of the maximum, average and standard derivation at traffic stations (A1 and A2) were the largest, while those in the park station (A5) were the lowest. Such phenomena could be a result of the site-specific primary sources of particulate matters. For instance, vehicle exhaust emission would be the primary source of particulate matter and nitrogen oxide at traffic stations; air pollutant emission from residential and commercial activities would be the primary source of particulate matter and nitrogen oxide at general stations; and atmospheric transport would be the primary trigger of particulate matter and nitrogen oxide at the park station because human activities would be less here. In other words, the driving force of air pollutants from vehicle transportation was stronger than that of the other human activities in Taipei City.

According to the highest values of the Kendall tau coefficients, the time lags of input variables were set as 1 h up to 4 h for five meteorological factors (rainfall, temperature, wind speed, wind direction and relative humidity) and eight air quality factors ($PM_{2.5}$, PM_{10}, O_3, NO_x, NO_2, NO, SO_2, CO). The SM-LSTM and DM-LSTM models were constructed for making regional multi-step-ahead (horizons $t+1$ up to $t+4$) air quality forecasts in Taipei City. The performance of these models was presented in terms of RMSE and G_{bench}.

4. Results and discussion

This study intended to explore and assess the predictability of the LSTM coupled with various deep learning algorithms on multiple outputs at different horizons for promoting the reliability and accuracy of regional air quality forecasts. The results and findings were presented and discussed in details in the order of model assessment, spatial stability of the models, temporal stability of the models, and summarization, shown as follows.

4.1. Performance of LSTM models for air quality forecasts in Taipei City

In this study, the number of input variables was 580 (N_{in} = (8

Table 1
Statistic indexes of observed $PM_{2.5}$, PM_{10} and NO_x concentrations at five air quality monitoring stations in Taipei City.

Season	Statistic index	Air Quality Monitoring Stations														
		A1			A2			A3			A4			A5		
		$PM_{2.5}$[a]	PM_{10}	NO_x	$PM_{2.5}$	PM_{10}	NO_x	$PM_{2.5}$	PM_{10}	NO_x	$PM_{2.5}$	PM_{10}	NO_x	$PM_{2.5}$	PM_{10}	NO_x
Spring	Maximum	377	457	325	358	401	287	259	350	216	278	366	237	147	183	113
	Mean	25	35	21	27	38	17	21	24	18	18	23	16	13	16	11
	Minimum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Standard Deviation	18	24	19	16	19	17	13	15	11	12	15	11	8	12	9
Summer	Maximum	226	283	188	215	257	182	155	204	137	167	212	153	88	117	71
	Mean	15	20	11	16	19	12	13	15	9	11	14	10	8	11	6
	Minimum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Standard Deviation	11	14	11	10	13	11	8	10	8	7	11	6	5	8	4
Autumn	Maximum	264	318	210	251	306	185	181	224	157	195	237	170	103	128	84
	Mean	18	23	15	19	21	13	15	18	11	13	20	15	9	11	7
	Minimum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Standard Deviation	13	15	10	11	16	10	9	13	8	8	14	9	6	9	8
Winter	Maximum	358	407	309	340	395	291	246	315	224	264	307	201	140	179	131
	Mean	24	30	22	26	27	23	20	24	18	17	25	14	12	17	10
	Minimum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Standard Deviation	17	21	14	15	18	12	12	16	13	11	14	9	8	14	11

[a] Units of $PM_{2.5}$, PM_{10} and NO_x concentrations are $\mu g/m^3$, $\mu g/m^3$, ppb, respectively.

(air quality factors)+5 (meteorological factors))x5 (air quality monitoring stations)x4 (time-lags)+5 (meteorological factors)x16 (meteorological monitoring stations)x4 (time-lags)) while the number of output variables was 15 ($= N_{out} = 5$ (air quality monitoring stations)x3 (air quality indexes, i.e. $PM_{2.5}$, PM_{10} and NO_x)). It was noted that these massive input variables were used for model construction in order to explore the suitability and effectiveness of the proposed DM-LSTM model incorporated with three learning algorithms (i.e., mini-batch gradient decent (MBGD) algorithm, dropout neuron algorithm and L2 regularization algorithm). The optimal numbers of hidden layers and neurons usually depend on the amount of datasets and system complexity. Through trial and error procedures, it is very likely to cause overfitting if the number of parameters exceeds 50% of training datasets while it is easy to reduce model generalizability if the number of parameters is less than 10% of training datasets. Taking the horizon t+4 for example, Table 2 presented the optimal parameters of the four LSTM models investigated in this study. It was noticed that: (1) the difference between SM-LSTM and DM-LSTM1 models was that the former was an SNN with only one hidden layer while the latter was a DNN with two or three hidden layers; (2) the difference between DM-LSTM1 and DM-LSTM2 models was that the learning algorithms adopted in the former and latter were the SGD and the MBGD, respectively; and (3) the difference between DM-LSTM2 and DM-LSTM3 models was that the latter used dropout neuron and L2 regularization algorithms while the former did not. The results shown in Table 2 indicated that: (1) the optimal number of neurons was 20 occurring at the minimal MSE value of 0.87 for the SM-LSTM model, while the optimal number of hidden layers was 2 occurring at the minimal MSE value of 0.81 for the DM-LSTM1 model; (2) the numbers of hidden layers were 1 and 2 for SM-LSTM and DM-LSTM1 models with their optimal parameters, respectively; and (3) the optimal mini-batch size of training datasets was 128 ($=2^7$) for the DM-LSTM2 model. It was noted that the comparison

Table 2
Parameters of the four LSTM models at horizon t+4.

Model	Parameters							MSE
	E_{max}	Neurons	Hidden layer	Learning rate α	Batch size	Dropout probability p	Coefficient λ	
SM-LSTM [a]	500	10	1	0.0005	N [e] = 35,064	/	/	1.13
		20						**0.87**
		30						1.19
		40						1.36
DM-LSTM1 [b]	500	20	**2**	0.0005	N = 35,064	/	/	**0.81**
			3					1.63
DM-LSTM2 [c]	500	20	2	0.0005	2^5 [f]	/	/	1.15
					2^6			1.08
					$\mathbf{2^7}$			**0.75**
					2^8			1.11
DM-LSTM3 [d]	500	20	2	0.0005	2^7	0.50	0.005	0.72

A value in bold denotes the optimal value of each parameter corresponding to the minimal value of Mean Square Error (MSE, is calculated with the normalized dataset) in training stages of each model.

[a] SM-LSTM denotes the Shallow Multi-output LSTM neural network with only 1 hidden layer, where the parameters are optimized by using the stochastic gradient decent algorithm (SGD).

[b] DM-LSTM1 denotes the Deep Multi-output LSTM neural network with 2 or 3 hidden layers, where the parameters are also optimized by using the SGD algorithm.

[c] DM-LSTM2 denotes the Deep Multi-output LSTM neural network with 2 hidden layers, where the parameters are optimized by using the mini-batch gradient decent algorithm (MBGD).

[d] DM-LSTM3 denotes the Deep Multi-output LSTM neural network with 2 hidden layers, where the parameters are optimized by the integration of the MBGD algorithm, dropout neuron algorithm and L2 regularization algorithm.

[e] N is the number of training dataset, where the GD algorithm is used to optimize model parameters.

[f] For identifying the mini-batch size, the MBGD is used to optimize model parameters.

analysis made between SM-LSTM and DM-LSTM1 models was to illustrate the difference in model performance between shallow and deep neural network models, while the comparison analysis made among three DM-LSTM models was to illustrate the contribution of various deep learning algorithms to overcoming the bottlenecks of instability and overfitting that occurred in regional multi-step-ahead air quality ($PM_{2.5}$, PM_{10}, and NO_x) forecasting. Besides, the dropout neuron algorithm with probability p (=0.5, in our case) could randomly delete some neurons from the hidden layer and reduce the number of model parameters during training for the avoidance of overfitting. As soon as the search process of the deep learning algorithm(s) obtained a set of the optimal weights in an iteration, the optimal number of hidden layers and the optimal number of hidden units (neurons) in each layer would be stored automatically while the hidden units together with their corresponding weights stored in the previous iteration would be removed then. In this study, the dropout neuron and the L2 regularization algorithms were responsible for optimizing the number of hidden units in each layer, while the MBGD algorithm was responsible for optimizing the number of hidden layers (Table 2). Additionally, the SGD algorithm failed to automatically determine the numbers of hidden layers and hidden units, and therefore would induce redundant hidden units. Moreover, the search process of the SGD algorithm was very time consuming because only one training dataset was used in each iteration. Therefore, the hybrid of the three deep learning algorithms (the MBGD, the dropout neuron, and the L2 regularization algorithms) would play a pivotal role in the automatic determination of a proper network size for DNN.

To show the merits of the proposed DM-LSTM models, an assessment was conducted on the results obtained from the training and testing stages of the four models at horizon t+4 regarding air quality ($PM_{2.5}$, PM_{10}, and NO_x) forecasts in Taipei City (Fig. 4). The comparison between SM-LSTM and DM-LSTM1 models showed that the final loss function value (0.81) of the DM-LSTM1 model was smaller than that (0.87) of the SM-LSTM model in the training stages while the final loss function value (1.14) of the DM-LSTM1 model was larger than that (0.98) of the SM-LSTM model in the testing stage. It indicated that overfitting occurred in the DM-LSTM1 model if the performance was good in the training stage but decreased significantly in the testing stage. In addition, the loss function values of the SM-LSTM model in both stages showed less fluctuation than those of the DM-LSTM1 model, which implied the DM-LSTM1 model would easily trigger forecast instability problem. The reason was that instability and overfitting bottlenecks would be easily induced by DNNs, for instance, the number of model parameters and model complexity increased as the number of hidden layers increased for the DM-LSTM1 model. The analyzed results indicated that the DNN model (i.e., DM-LSTM1) required more auxiliary deep learning techniques to handle its instability and overfitting problems. Next, the comparison between DM-LSTM1 and DM-LSTM2 models showed that the loss function values of the DM-LSTM2 model fluctuated less and were smaller, which implied the DM-LSTM2 model would overcome the forecast instability owing to the utilization of the MBGD algorithm in the training stage. Then, the comparison between DM-LSTM2 and DM-LSTM3 models indicated that the difference (0.09 = 0.81−0.72) of the final loss function values between the training and testing stages of the DM-LSTM3 model was significantly smaller than that (0.17 = 0.92−0.75) of the DM-LSTM2 model. It demonstrated that the DM-LSTM3 model overcame the overfitting bottleneck occurring in the deep learning neural network because of the utilization of the dropout neuron and L2 regularization algorithms during the training stage.

These comparative results demonstrated that the proposed DM-LSTM3 model with three deep learning algorithms not only produced the smallest loss function values as well as the most stable loss function curve but also effectively overcame the instability and overfitting shortcomings for regional multi-step-ahead air quality forecasting. Such achievement made by the DM-LSTM3 model could be owing to the reasons that the dropout neuron and L2 regularization algorithms improved forecast accuracy from the perspective of tackling the overfitting bottleneck while the MBGD algorithms improved forecast accuracy from the perspective of overcoming the instability bottleneck.

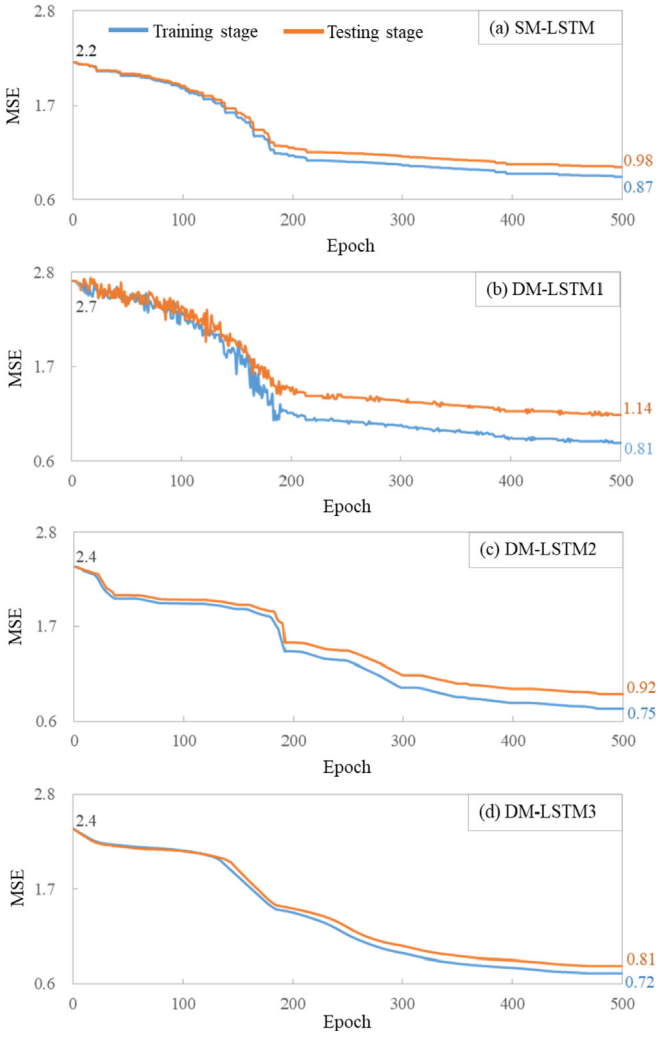

Fig. 4. Mean Squared Error (MSE) values of LSTM models in training and testing stages at horizon t+4 (the value of MSE is calculated with the normalized dataset).

4.2. Spatial stability of LSTM models

The spatial stability of these constructed models was assessed. Taking the RMSE values in testing stages as an example, Fig. 5 showed the improvement rates of the DM-LSTM3 model over the SM-LSTM model in regional and site-specific multi-step-ahead forecasting of $PM_{2.5}$, PM_{10} and NO_x concentrations at horizons t+1 up to t+4 at five monitoring stations in Taipei City, respectively. The results clearly showed the following findings.

1) The DM-LSTM3 model integrated with three deep learning algorithms produced the best performance not only on multi-

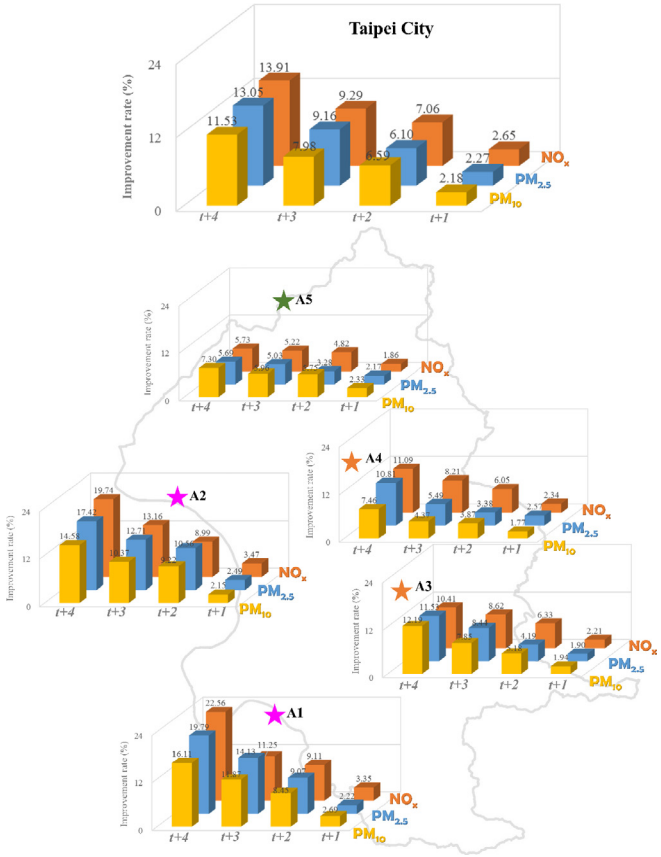

Fig. 5. Improvement rates (*RMSE (μg/m^3)) of the DM-LSTM3 model over the SM-LSTM model in the testing stages of the multi-step-ahead forecasting models for $PM_{2.5}$, PM_{10} and NO_x, respectively.*Improvement rate in RMSE $= \frac{(RMSE\{SM\_LSTM\} - RMSE\{DM\_LSTM3\})}{RMSE\{SM\_LSTM\}} \times 100\%$

step-ahead forecasting but also on multi-output forecasting. For instance, for $PM_{2.5}$, PM_{10} and NO_x at horizon t+4, the values of RMSE were 9.31 μg/m^3, 10.62 μg/m^3 and 12.17 ppb accordingly while the G_{bench} values were 0.83, 0.81 and 0.82 accordingly (Table 3). It was noted that the improvement rates in RMSE significantly increased at time steps t+3 and t +4 (Fig. 5). The results demonstrated that the proposed M-LSTM3 model could provide reliable and accurate regional multi-step-ahead air quality forecasts because it adequately considered the underlying non-linear spatial relationships among five air quality monitoring stations to effectively adjust synaptic weights.

2) For each air quality monitoring station, the DM-LSTM3 model also produced the best testing performance of all the cases regarding the improvement rates in RMSE. It appeared that the DM-LSTM3 model produced much smaller RMSE values than the SM-LSTM model did in the testing stage. The DM-LSTM3 model performed significantly better at traffic stations (A1 and A2) while slightly better at general stations (A3, and A4) and the park station (A5), as compared to the SM-LSTM model. In addition, it was an interesting finding that the improvement rates in RMSE significantly increased from t+3 to t+4 at all stations. Taking horizon t+4 for example, the improvement rates in RMSE for PM_{10}, $PM_{2.5}$ and NO_x forecasts reached 16.11%, 19.79% and 22.56% at Station A1, respectively, but reduced to 7.30%, 5.69% and 5.73% at Station A5, respectively. In other words, the proposed DM-LSTM3 model could make more stable and accurate multi-step-ahead forecasts through identifying the heterogeneities among different air quality monitoring stations. This could be because the correlation between air quality concentrations and traffic stations stemmed not only from traffic volumes but also from meteorological factors (e.g., rainfall and wind speed), while the correlation air quality concentrations and between general stations stemmed only from meteorological factors in the perspective of spatial relationship. Regarding the Kendall tau correlation for $PM_{2.5}$, PM_{10} and NO_x, their correlation values with traffic stations (coefficients = 0.87, 0.82 and 0.92, respectively) were higher than those with the other stations (average coefficients = 0.73, 0.65 and 0.55, respectively). Therefore, the DM-LSTM3 model performed the best at traffic stations. The results indicated that the deep learning multi-output LSTM neural network model utilized the architecture of multiple ($\geq$ 2) hidden layers to capture the inherent features of data layer-by-layer without prior knowledge and thus performed well in regional multi-step-ahead air quality forecasting.

From the perspective of spatial stability, the DM-LSTM model was very beneficial to regional air quality forecasting since the proposed deep learning-based multi-output data-driven model hybriding three deep learning algorithms could enhance model reliability and forecast accuracy.

4.3. Temporal stability of LSTM models

Next, the temporal stability of these constructed models was evaluated. Table 4 presented the improvement rates of the DM-LSTM3 model over the SM-LSTM model in terms of RMSE for regional multi-step-ahead air quality ($PM_{2.5}$, PM_{10}, and NO_x) forecasts in four seasons. Several findings were found and expressed as follows. At a regional scale (Taipei City), the DM-LSTM3 model had the best performance in four seasons. It was noted that for horizons

Table 3
Performance of (RMSE & G_{bench}) in the testing stages of the multi-step-ahead forecasting models for $PM_{2.5}$, PM_{10} and NO_x at horizons from $t + 1$ to $t + 4$ in Taipei City (the DM-LSTM3 model in comparison with the SM-LSTM model).

Indicator	Horizon	$PM_{2.5}$		PM_{10}		NO_x	
		SM-LSTM	DM-LSTM3	SM-LSTM	DM-LSTM3	SM-LSTM	DM-LSTM3
RMSE[a]	t+1	4.59	4.49	5.55	5.43	6.75	6.57
	t+2	5.74	5.40	7.08	6.62	8.05	7.48
	t+3	7.37	6.70	10.05	9.26	11.07	10.04
	t+4	10.70	9.31	11.99	10.62	14.13	12.16
G_{bench}	t+1	0.93	0.97	0.92	0.95	0.91	0.95
	t+2	0.88	0.92	0.87	0.91	0.86	0.90
	t+3	0.82	0.87	0.81	0.85	0.82	0.87
	t+4	0.72	0.83	0.73	0.81	0.72	0.82

[a] Units of RMSE for $PM_{2.5}$, PM_{10} and NO_x concentrations are μg/m^3, μg/m^3, ppb, respectively.

Table 4
Improvement rates of seasonal performance (Root-Mean-Square-Error, RMSE) in the testing stages of the multi-step-ahead forecasting models for $PM_{2.5}$, PM_{10} and NO_x at horizons from $t+1$ up to $t+4$ in Taipei City (the DM-LSTM3 model in comparison with the SM-LSTM model).

Season	Horizon	Improvement rate (%)[a] of RMSE[b]		
		$PM_{2.5}$	PM_{10}	NO_x
Spring	t+1	2.31	1.77	1.86
	t+2	8.85	3.87	4.82
	t+3	9.12	4.37	5.22
	t+4	14.24	9.08	5.73
Summer	t+1	2.65	2.69	3.47
	t+2	12.57	8.45	8.99
	t+3	18.22	13.92	13.16
	t+4	22.88	16.11	19.74
Autumn	t+1	2.49	2.08	3.35
	t+2	10.56	9.29	9.11
	t+3	12.71	10.43	11.25
	t+4	17.42	15.47	17.76
Winter	t+1	1.76	2.16	2.34
	t+2	8.75	6.52	6.05
	t+3	8.68	7.90	8.21
	t+4	10.18	11.19	11.09

[a] Improvement rate in RMSE $= \frac{(\mathrm{RMSE\{SM\_LSTM\}} - \mathrm{RMSE\{DM\_LSTM3\}})}{\mathrm{RMSE\{SM\_LSTM\}}} \times 100\%$

[b] Units of RMSE for $PM_{2.5}$, PM_{10} and NO_x concentrations are $\mu g/m^3$, $\mu g/m^3$, ppb, respectively.

t+3 and t+4, the DM-LSTM3 model performed much more excellently in summer and autumn while slightly better in spring and winter, as compared to the SM-LSTM model. In addition, it was an interesting finding that the improvement rates in RMSE significantly increased from horizon t+1 up to horizon t+4 in all seasons. Taking horizon t+4 for example, the improvement rates in RMSE for $PM_{2.5}$, PM_{10} and NO_x forecasting reached 22.88%, 16.11% and 19.74% in summer, respectively, but reduced to (only) 10.18%, 11.19% and 11.09% in winter, respectively. The reason was that the meteorological relationships with air quality in summer and autumn (average Kendall tau coefficient for meteorological factors = 0.75) was stronger than those in spring and winter (average Kendall tau coefficient for meteorological factors = 0.62). Therefore, the stronger correlation between air quality monitoring stations was considered as a trigger to enhance the forecast performance of the DM-LSTM3 model. In short, these findings were also very beneficial to data-driven modellers because the proposed deep learning-based multi-output LSTM neural network model could provide accurate and reliable regional multi-step-ahead air quality forecasts in four seasons, according to the temporal stability comparison between SM-LSTM and DM-LSTM3 models.

Finally, to clearly distinguish the predictability between SM-LSTM and DM-LSTM3 models, an air pollution event with its maximal $PM_{2.5}$, PM_{10} and NO_x concentrations higher than 150 $\mu g/m^3$, 200 $\mu g/m^3$ and 150 ppb, respectively, at a traffic station (A2, San-Chong) was selected to test both models through assessing the goodness-of-fit between observations and forecasts at horizon t+4, as shown in Fig. 6. The results revealed that the DM-LSTM3 model was able to well forecast air quality at horizon t+4, whereas the SM-LSTM model failed to forecast accurately due to obvious time-lag phenomena and larger gaps between observations and forecasts. It appeared that the developed DM-LSTM model could trace the trails of air quality events, significantly mitigate time-lag effects, as well as make much accurate and reliable regional multi-step-ahead air quality forecasts.

In summary, regional air quality frequently interacts with intensive human activities, traffic loads and commercial trading in cities with fast urban development like Taipei City. In this study, air quality monitoring stations A1 and A2 are traffic stations representative of traffic loads for monitoring the primary air pollutant mechanism. Air quality monitoring stations A3 and A4 are general stations representative of human activities and commercial trading for monitoring the secondary air pollutant mechanism. Air quality monitoring station A5 is a park station representative of natural situations with less human intervention. From the perspective of monitoring functions and spatial distribution, these five air quality monitoring stations are typical and representative for regional air quality of Taipei City. Besides, epidemiological research pointed out that exposure to PM2.5, PM10, and NOx could result in air pollution-related diseases, which implied the importance of air quality forecasting in environmental management decision-making and prevention from serious air pollution-related accidents. However, traditional methods were used mainly to establish the site-specific data-driven model for each individual air quality monitoring station. It would generally be more promising for multi-output data-driven models to extract the underlying nonlinear interrelationship among output variables and improve regional forecast accuracy. Nevertheless, multi-output air quality forecasting increased modelling difficulty when SNN models were implemented. Bearing this in mind as a motivation, the innovative nature of this study was indebted to: the hybrid of the LSTM and three state-of-the-art deep learning algorithms for achieving accurate regional forecasts at different horizons through overcoming model instability and overfitting; and its application for the first time to regional multi-step-ahead air quality forecasting. In comparison with similar studies, such as the single-output data-driven techniques (Prasad et al., 2016; Yeganeh et al., 2018; Nieto et al., 2018) and multi-output data-driven techniques (Nguyen et al., 2012; Li et al., 2016), the main findings of this study were in the best interests of the goal to make the Earth a better place to live and were explored on the grounds that: (1) the proposed DM-LSTM model could effectively increase the accuracy of regional multi-step-ahead air-quality forecasts through tackling error accumulation and propagation commonly encountered in regional forecasting; and (2) the proposed DM-LSTM model could be effectively applied not only to modelling the heterogeneities in different air pollutant-generating mechanisms (e.g., primary and secondary mechanisms, and natural situations) but also to mapping the heterogeneous air pollutants onto different seasons by utilizing DNNs to capture the inherent features of the data layer-by-layer without prior knowledge for describing the potentially non-linear interrelationships among $PM_{2.5}$, PM_{10} and NO_x monitoring stations. Consequently, the proposed DM-LSTM model equipped with three deep learning algorithms proved to be spatio-temporally stable and was considered the most suitable for regional air quality forecasting at different lead times.

5. Conclusion

People across development cities like Taipei City undergo a great possibility of exposure to high-level invasion of air pollutants. Thus, accurate and reliable regional multi-step-ahead air quality forecasting is very crucial and beneficial to reduce health risks caused by ambient air pollution. In this paper, a deep learning-based multi-output LSTM neural network model (DM-LSTM) equipped with three deep learning algorithms was first proposed for regional multi-step-ahead air quality ($PM_{2.5}$, PM_{10} and NO_x) forecasting. Its capability of efficient learning and accurate forecasting was tested and verified at five air quality monitoring stations in Taipei City. The Shallow Multi-output LSTM model (SM-LSTM) was implemented for comparative analysis.

The results of regional air quality ($PM_{2.5}$, $PM_{10,}$ and NO_x) forecasts demonstrated that the proposed DM-LSTM model performed prominently than the SM-LSTM model in multi-step-ahead

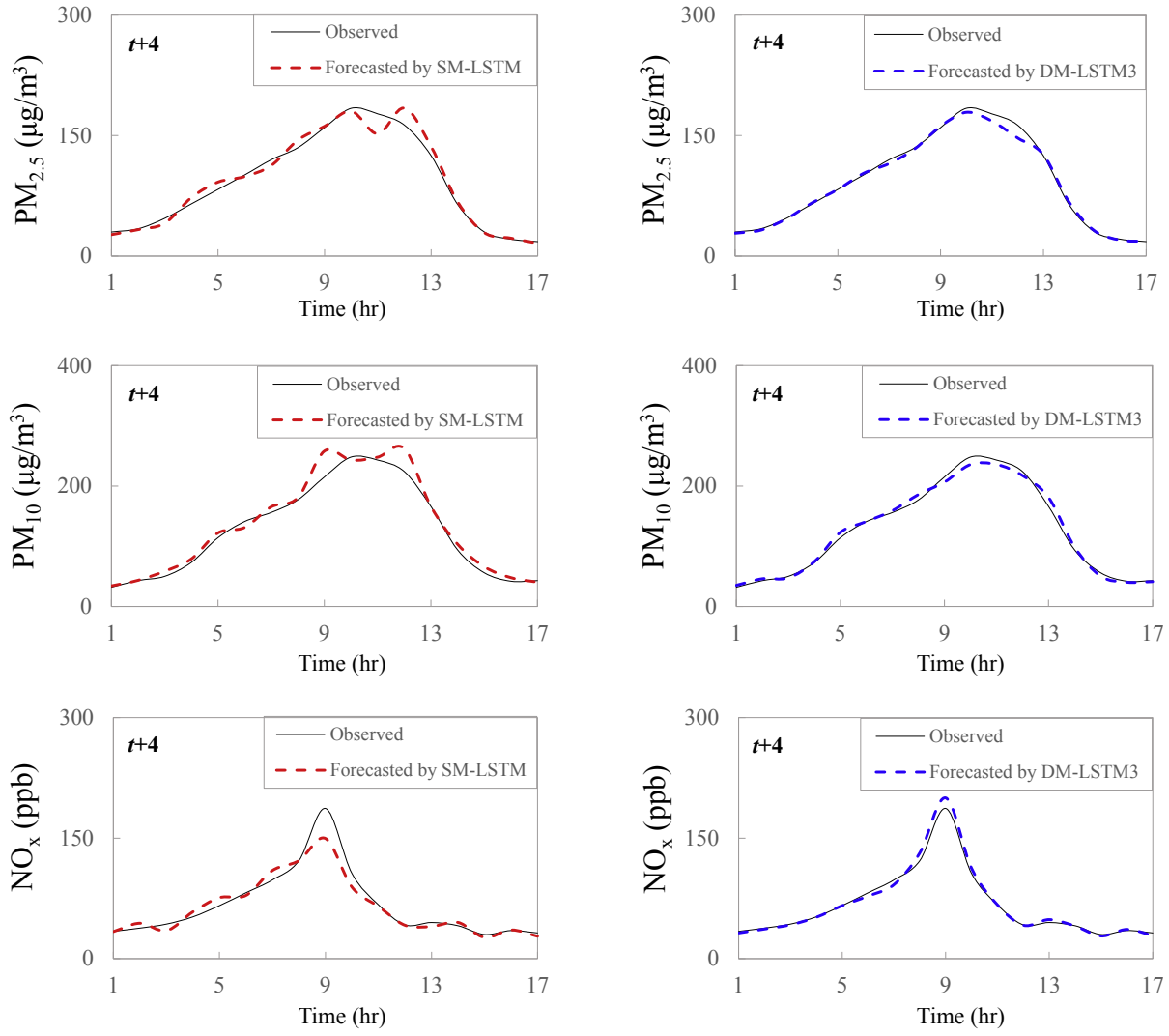

Fig. 6. $PM_{2.5}$, PM_{10} and NO_x forecast results at horizon $t + 4$ at traffic station A2 (San-Chong) using SM-LSTM and DM-LSTM3 models. The highest-peaks of $PM_{2.5}$, PM_{10} and NO_x concentrations in the testing stages exceed 150 μg/m^3, 200 μg/m^3 and 150 ppb, respectively.

forecasting for all the cases, with smaller RMSE (improvement rates ranged from 2.18% to 13.91%, Fig. 5) and larger G_{bench} values (improvement rates ranged from 3.26% to 15.28%, Table 3) for the Taipei City. It showed that the proposed DM-LSTM model that adequately extracted underlying non-linear spatial relationships among five air quality monitoring stations could effectively adjust synaptic weights and provide reliable and accurate regional multi-step-ahead forecasts on $PM_{2.5}$, $PM_{10,}$ and NO_x. When assessing the regional air quality forecast models established for Taipei City, the proposed DM-LSTM model could significantly mitigate time-lag phenomena and solve the overfitting problem. In contrast, the SM-LSTM model produced an inferior performance at all horizons, especially t+3 and t+4, in both training and testing stages. It implied that the SM-LSTM model demanded for more sophisticated techniques, such as deep learning algorithms with two or more hidden layers, to improve model stability and generalizability at spatio-temporal scales. The developed deep learning-based multi-output LSTM neural network model (DM-LSTM) could effectively capture the heterogeneities in different air pollutant-generating mechanisms (e.g., primary and secondary mechanisms, and natural situations) and adequately map the heterogeneous air pollutants onto different seasons. Therefore, the DM-LSTM model incorporated with three deep learning algorithms could provide early forecasting and warming on regional air quality for reducing health risks associated with outdoor activities.

Acknowledgment

This study is financially supported by the Ministry of Science and Technology, Taiwan, ROC (MOST: 106-3114-M-002-001-A and 106-2811-B-002 -087-) and the China Postdoctoral Science Foundation (No. 2017M620336).

Appendix A

Original Long Short-Term Memory (LSTM) unit

The LSTM unit is composed of six parts, including input block, three gates (input, forget and output gates), self-looped cell as well as output block. The equations shown below describe how an LSTM unit is updated at every time step t.

(1) The input block is used to produce current memory information ($\tilde{C}_t$) by combining the model input (x_t) with the output of the previous state (h_{t-1}).

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (1)$$

where tanh is a hyperbolic tangent function. W_c is the weight for the input of the current state in the input block. U_c is the weight for the output of the previous state in the input block. b_c is the bias in the input block.

(2) The input gate (i_t) is able to decide what information to add to the current cell state by learning from the output of the previous state (h_{t-1}) and the input of the current state (x_t).

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

where σ is a sigmoid transfer function. W_i is the weight for the input of the current state in the input gate. U_i is the weight for the output of the previous state in the input gate. b_i is the bias in the input gate.

(3) The forget gate (f_t) can decide what information to remove from the current cell state by learning from the output of the previous state (h_{t-1}) and the input of the current state (x_t).

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (3)$$

where W_f is the weight for the input of the current state in the forget gate. U_f is the weight for the output of the previous state in the forget gate. b_f is the bias in the forget gate.

(4) The self-looped cell (C_t) can create an update for the previous self-looped cell state (C_{t-1}) by combining the information of the input and forget gates with current input block ($\tilde{C}_t$).

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (4)$$

(5) The output gate (o_t) can decide the output of the self-recurrent cell. The tanh is utilized to transform the self-looped cell state (C_t) to ensure that the value falls within [-1, 1] and the transformed result is multiplied by the value of the output gate, which produces the current output state (h_t).

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_t + b_o) \quad (5a)$$

$$h_t = o_t * \tanh(C_t) \quad (5b)$$

where W_o is the weight for the input of the current state in the output gate. U_o is the weight for the output of the previous state in the output gate. V_o is the weight for the self-recurrent cell state in the output gate. b_o is the bias in the output gate.

(6) The output block is used to compute the output of the LSTM unit, which is considered as the algebraic sum of the output gate.

$$y(t) = W_y h_t + b_y \quad (6)$$

where $y(t)$ is the output of the LSTM unit. W_o is the weight for the current output state. b_y is the bias in the output block.

References

Afzali, A., Rashid, M., Afzali, M., Younesi, V., 2017. Prediction of air pollutants concentrations from multiple sources using AERMOD coupled with WRF prognostic model. J. Clean. Prod. 166, 1216–1225.

Apte, J.S., Marshall, J.D., Cohen, A.J., Brauer, M., 2015. Addressing global mortality from ambient $PM_{2.5}$. Environ. Sci. Technol. 49 (13), 8057–8066.

Ausati, S., Amanollahi, J., 2016. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting $PM_{2.5}$. Atmos. Environ. 142, 465–474.

Baldi, P., Sadowski, P., 2014. The dropout learning algorithm. Artif. Intell. 210, 78–122.

Ballesteros, M., Dyer, C., Goldberg, Y., Smith, N.A., 2017. Greedy transition-based dependency parsing with stack LSTMs. Comput. Ling. 43 (2), 311–347.

Bilgic, B., Chatnuntawech, I., Fan, A.P., Setsompop, K., Cauley, S.F., Wald, L.L., et al., 2014. Fast image reconstruction with L2-regularization. J. Magn. Reson. Imag. 40 (1), 181–191.

Chang, F.J., Kao, L.S., Kuo, Y.M., Liu, C.W., 2010. Artificial neural networks for estimating regional arsenic concentrations in a Blackfoot disease area in Taiwan. J. Hydrol. 388, 65–76.

Chang, F.J., Chen, P.A., Chang, L.C., Tsai, Y.H., 2016. Estimating spatio-temporal dynamics of stream total phosphate concentration by soft computing techniques. Sci. Total Environ. 562, 228–236.

Chang, F.J., Tsai, Y.H., Chen, P.A., Coynel, A., Vachaud, G., 2015. Modeling water quality in an urban river using hydrological factors—data driven approaches. J. Environ. Manag. 151, 87–96.

Corani, G., Scanagatta, M., 2016. Air pollution prediction via multi-label classification. Environ. Model. Softw. 80, 259–264.

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.

Gao, M., Yin, L., Ning, J., 2018. Artificial neural network model for ozone concentration estimation and Monte Carlo analysis. Atmos. Environ. 184, 129–139.

Gong, B., Ordieres, M.J., 2016. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong. Environ. Model. Softw. 84, 290–303.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2017. LSTM: a search space odyssey. IEEE Trans. Neural. Netw. Learn. Syst. 28 (10), 2222–2232.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. Comput. Sci. 3 (4), 212–223.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets, and problem solutions. Int. J. Uncertain. Fuzziness Knowledge-Based Syst. 6 (2), 107–116.

Jiang, B., Xia, D., Zhang, X., 2018. A multicomponent kinetic model established for investigation on atmospheric new particle formation mechanism in H_2SO_4-HNO_3-NH_3-VOC system. Sci. Total Environ. 616, 616–617.

Kabán, A., 2013. Fractional norm regularization: learning with very few relevant features. IEEE Trans. Neural. Netw. Learn. Syst. 24 (6), 953–963.

Karambelas, A., Holloway, T., Kiesewetter, G., Heyes, C., 2018. Constraining the uncertainty in emissions over India with a regional air quality model evaluation. Atmos. Environ. 174, 194–203.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Lauret, P., Heymes, F., Aprin, L., Johannet, A., 2016. Atmospheric dispersion modeling using Artificial Neural Network based cellular automata. Environ. Model. Softw. 85, 56–69.

Lin, B., Zhu, J., 2018. Changes in urban air quality during urbanization in China. J. Clean. Prod. 188, 312–321.

Li, H., You, S., Zhang, H., Zheng, W., Lee, W.L., Ye, T., Zou, L., 2018a. Analyzing the impact of heating emissions on air quality index based on principal component regression. J. Clean. Prod. 171, 1577–1592.

Li, N., Chen, J.P., Tsai, C.I., et al., 2017. Potential impacts of electric vehicles on air quality in Taiwan. Sci. Total Environ. 566, 919–928.

Li, X., Peng, L., Hu, Y., Shao, J., Chi, T., 2016. Deep learning architecture for air quality predictions. Environ. Sci. Pollut. Res. 23 (22), 22408–22417.

Li, L., Lei, Y., Wu, S., Huang, Z., Luo, J., Wang, Y., Yan, D., 2018b. Evaluation of future energy consumption on $PM_{2.5}$ emissions and public health economic loss in Beijing. J. Clean. Prod. 187, 1115–1128.

Liu, M., Bi, J., Ma, Z., 2017. Visibility-based $PM_{2.5}$ concentrations in China: 1957-1964 and 1973-2014. Environ. Sci. Technol. 51, 13161–13169.

Maidment, D.R., 1993. Handbook of Hydrology. McGraw-Hill, New York.

Nakama, T., 2009. Theoretical analysis of batch and on-line training for gradient descent learning in neural networks. Neurocomputing 73 (1), 151–159.

Nguyen, V.A., Starzyk, J.A., Goh, W.B., Jachyra, D., 2012. Neural network structure for spatio-temporal long-term memory. IEEE Trans. Neural Netw. Learn Syst. 23 (6), 971–983.

Ni, X.Y., Huang, H., Du, W.P., 2017. Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. Atmos. Environ.

150, 146–161.
Nielsen, M.A., 2015. Neural Networks and Deep Learning. Determination Press.
Nieto, P.J.G., García-Gonzalo, E., Sánchez, A.B., Miranda, A.A.R., 2017. Air quality modeling using the PSO-SVM-based approach, MLP neural network, and M5 model tree in the metropolitan area of Oviedo (Northern Spain). Environ. Model. Assess. 23 (4), 1–19.
Nieto, P.J.G., Lasheras, F.S., García-Gonzalo, E., Juez, F.J.D.C., 2018. PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. Sci. Total Environ. 621, 753–761.
Pisoni, E., Albrecht, D., Mara, T.A., Rosati, R., Tarantola, S., Thunis, P., 2018. Application of uncertainty and sensitivity analysis to the air quality sherpa modelling tool. Atmos. Environ. 183, 84–93.
Pournazeri, S., Tan, S., Schulte, N., Jing, Q., Venkatram, A., 2014. A computationally efficient model for estimating background concentrations of NO_x, NO_2, and O_3. Environ. Model. Softw. 52, 19–37.
Prasad, K., Gorai, A.K., Goyal, P., 2016. Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. Atmos. Environ. 128, 246–262.
Qian, Q., Jin, R., Yi, J., Zhang, L., Zhu, S., 2015. Efficient distance metric learning by adaptive sampling and mini-batch stochastic gradient descent (SGD). Mach. Learn. 99 (3), 353–372.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1985. Learning representations by back-propagating errors. Nature 323, 533–536.
Reggente, M., Peters, J., Theunis, J., Van Poppel, M., Rademaker, M., Kumar, P., De Baets, B., 2014. Prediction of ultrafine particle number concentrations in urban environments by means of Gaussian process regression based on measurements of oxides of nitrogen. Environ. Model. Softw. 61, 135–150.
Reisen, V.A., Sarnaglia, A.J.Q., Reis Jr., N.C., Lévy-Leduc, C., Santos, J.M., 2014. Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility. Environ. Model. Softw. 51, 286–295.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Taghavifar, H., Taghavifar, H., Mardani, A., Mohebbi, A., Khalilarya, S., Jafarmadar, S., 2016. Appraisal of artificial neural networks to the emission analysis and prediction of CO_2, soot, and NOx of n-heptane fueled engine. J. Clean. Prod. 112, 1729–1739.
Taylan, O., 2017. Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality. Atmos. Environ. 150, 356–365.
Van Fan, Y., Perry, S., Klemeš, J.J., Lee, C.T., 2018. A review on air emissions assessment: Transportation. J. Clean. Prod. 194, 673–684.
Vijayaraghavan, K., Cho, S., Morris, R., Spink, D., Jung, J., Pauls, R., et al., 2016. Photochemical model evaluation of the ground-level ozone impacts on ambient air quality and vegetation health in the Alberta oil sands region: using present and future emission scenarios. Atmos. Environ. 141, 209–218.
Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ. 409 (7), 1266–1276.
Wakeel, M., Yang, S., Chen, B., Hayat, T., Alsaedi, A., Ahmad, B., 2017. Network perspective of embodied PM2. 5-A case study. J. Clean. Prod. 142, 3322–3331.
Wang, Y., Sun, M., Yang, X., Yuan, X., 2016. Public awareness and willingness to pay for tackling smog pollution in China: a case study. J. Clean. Prod. 112, 1627–1634.
Wang, J., Cao, Z., 2017. Chinese text sentiment analysis using LSTM network based on L2 and Nadam. In: Communication Technology (ICCT), 2017 IEEE 17th International Conference, pp. 1891–1895.
Wei, X., Lin, H., Yang, L., Yu, Y., 2017. A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. Information 8 (3), 92–107.
Wu, L., Li, N., Yang, Y., 2018a. Prediction of air quality indicators for the Beijing-Tianjin-Hebei region. J. Clean. Prod. 196, 682–687.
Wu, J., Zheng, H., Zhe, F., Xie, W., Song, J., 2018b. Study on the relationship between urbanization and fine particulate matter ($PM_{2.5}$) concentration and its implication in China. J. Clean. Prod. 182, 872–882.
Yang, G., Huang, J., Li, X., 2018. Mining sequential patterns of PM2.5 pollution in three zones in China. J. Clean. Prod. 170, 388–398.
Yang, Y., Christakos, G., 2015. Spatiotemporal characterization of ambient $PM_{2.5}$ concentrations in Shandong province (China). Environ. Sci. Technol. 49 (22), 13431–13438.
Yeganeh, B., Hewson, M.G., Clifford, S., Tavassoli, A., Knibbs, L.D., Morawska, L., 2018. Estimating the spatiotemporal variation of NO_2 concentration using an adaptive neuro-fuzzy inference system. Environ. Model. Softw. 100, 222–235.
Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J., 2018a. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. J. Hydrol. 561, 918–929.
Zhang, D., Lindholm, G., Ratnaweera, H., 2018b. Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring. J. Hydrol. 556, 409–418.