Spatial estimation of outdoor NO_2 levels in Central London using deep neural networks and a wavelet decomposition technique

Sheen Mclean Cabaneros[a,*], John Kaiser Calautit[b], Ben Hughes[a]

[a] *Department of Mechanical and Aerospace Engineering, University of Strathclyde, Glasgow G1 1XQ*
[b] *Department of Architecture and Built Environment, University of Nottingham, Nottingham NG7 2RD*

Abstract

Outdoor air pollution remains a major environmental threat to the public, especially those who reside in highly urbanised areas. Recent studies have revealed the effectiveness of early-warning mechanisms that enable the public reduce their exposure to air pollutants. This highlights the need for accurate air quality forecasts. However, the air quality of many areas in developing and highly urbanised countries remains unmonitored. Hence, a novel spatiotemporal interpolation modelling approach based on deep learning and wavelet pre-processing technique was proposed in this paper. In more detail, Long Short-term Memory (LSTM) models and Discrete Wavelet Transformation (DWT) were utilised to model the spatial variability of hourly NO_2 levels at six urban sites in Central London, the United Kingdom. The models were trained using only the NO_2 concentration data from the neighbouring sites. Benchmark models such as plain LSTM and feed-forward neural network models were also developed to evaluate the effectiveness of the proposed model. The proposed wavelet-based models were found to provide superior forecasting results, explaining 77% to 93% of the variability of the actual NO_2 concentrations at most sites. The overall results reveal the very promising potential of the proposed models for the spatial characterisation of air pollution.

Keywords: Air pollution forecasting, Artificial neural networks, Wavelet Decomposition, Long short-term memory Units, Deep learning, Spatial interpolation

1. Introduction

Outdoor air pollution is considered as a major environmental concern attracting special attention from both scientific and decision-making communities. Poor air quality has been linked to several respiratory and cardiovascular illnesses and premature deaths (WHO, 2016), as well as major economic consequences due to health expenditures and productivity losses (OECD, 2016). As such, the problem of poor air quality clearly needs to be addressed to mitigate its adverse impacts.

One way to manage outdoor air pollution is through legislative standards imposed at local and national levels. For instance, the Ambient Air Quality Directive (2008/50/EC) which is established by the European Union (EU) sets up legal limits and target levels for several outdoor air pollutants (European Environmental Agency, 2018). Furthermore, air

[*] Corresponding author

quality forecasts can assist environmental and urban city planners in making well-informed decisions in managing air pollution (Baklanov et al., 2007; Gers et al., 2001). Forecasting tools are also used as early-warning mechanisms to aid the vulnerable members of the public to reduce their exposure to pollutants especially during peak pollution episodes (McLaren and Williams, 2015; Chen et al., 2018). As such, the development of air pollution forecasting tools remains an active topic of research (Cabaneros et al., 2019; Conti et al., 2017).

Air pollution modelling approaches are usually classified as either deterministic or data-driven. Deterministic approaches such as the Urban Airshed Model (UAM) (Chang and Cardelino, 2000) and Weather Research and Forecasting Model with Chemistry (WRF/Chem) (Chuang et al., 2011) apply physics-based principles to describe the formation, generation and dispersion of air pollutants in the ambient environment (Jacobson, 1997). However, deterministic models rely on default parameters and demand large computational resources limiting their use to many case studies (Dutot et al., 2007; National Research Council, 2007). Alternatively, data-driven approaches apply statistical techniques to estimate the input-output dynamics between air pollution levels and known explanatory variables only through past data. Popular methods employed in recent years include the Multiple Linear Regression (MLR) (Ng and Awang, 2018), Artificial Neural Network (ANN) (Dotse et al., 2018), Support Vector Machine (SVM) (Liu et al., 2017), and hybrid models (Franceschi et al., 2018).

ANN models have recently become popular in air pollution modelling applications because of their ability to approximate non-linear relationships (Cabaneros et al., 2019; Gardner and Dorling, 1998; Shahraiyni and Sodoudi, 2016). Since the interaction between meteorological parameters and most air pollutant species is complex and highly nonlinear (Colls, 2001), ANN models have been shown to outperform traditional linear statistical methods (Shahraiyni and Sodoudi, 2016; Alam and McNabola, 2015).

However, the use of ANN models in areas with a lack or absence of monitoring stations may be challenging. In such cases, spatial interpolation and extrapolation techniques in which available measurements from neighbouring stations can be employed to estimate the air pollutant concentration levels in unmonitored stations. For instance, Alimissis et al. (2018) spatially estimated several roadside pollution levels at several locations using feed-forward ANN and MLR models. A similar methodology was employed by Tzanis et al. (2019) to spatially estimate fine particle levels at several sites. However, the use of plain architectures may limit the performance of ANN models given the extremely complex dynamics between meteorological and air pollutant variables (Gardner and Dorling, 1998). Plain ANN models also display difficulties when estimating air pollution levels with high variables at a local scale (Siwek and Osowski, 2012).

To address the variability issue, the use of pre-processing techniques have been suggested in the past (Cabaneros et al., 2019). For instance, works dealing with the integration of wavelets and ANN models are well documented in the literature: wavelet-based ANN models (Feng et al., 2015; Bai et al., 2016), wavelet-based SVM models (Osowski and Garanty, 2007), an ensemble of plain and wavelet-based predictors (Siwek and Osowski, 2012). However, the said works did not extend the hybrid models for the spatial interpolation of air pollution levels. Moreover, most works do not take into account the long-term temporal dependencies of air pollution data. That is, the previous models do not capture the sequential temporal features of the input data limiting their performance (Freeman et al., 2018).

Deep learning models are recent additions to the family of ANN models which are capable of revealing temporal pattern from past information. Approaches based on deep learning have been shown to exhibit superior performance over standard ANN types (Cabaneros et al., 2019; Freeman et al., 2018; Li et al., 2017). For example, Wu and Lin (2019) proposed an approach comprising of a two-stage decomposition stage and the use of Long Short-term Memory (LSTM) models to forecast several pollutant levels at several sites. Li and Zhu (2018) developed a hybrid approach based on Extreme Learning Machine (ELM) and Imperialist Competitive Algorithm (ICA) to forecast pollutant concentrations at several sites. However, the said works do not extend their models in the forecasting of pollutant levels in unmonitored locations.

Limited studies have applied deep learning approaches to the spatiotemporal forecasting of air pollution levels. For instance, Zhao et al. (2019) employed a fully-connected LSTM model estimated the spatiotemporal variability of $PM_{2.5}$ levels a single site. Ma et al. (2019) combined Inverse Distance Weighting (IDW) technique and bi-directional LSTM model for the spatiotemporal forecasting of $PM_{2.5}$ levels over several locations. Wang and Song (2018) developed a deep spatiotemporal ensemble model based on LSTM at several locations. However, the said works do not incorporate any pre-processing techniques that can improve the performance of their models.

To address the limitations of previous works, this paper presents a novel approach in modelling the spatiotemporal variations of hourly NO_2 levels in Central London 1 hour in advance. The main contributions of this paper are two-fold. Firstly, this paper tests the ability of wavelet-based LSTM models in forecasting air pollution levels. This is significant as the use of wavelet decomposition with sophisticated forms of ANN models in the context of spatial interpolation of air pollution concentration has been very limited (Cabaneros et al., 2019). The results of this work also provide new insights regarding the modelling of NO_2 levels in other locations sharing characteristics with the monitoring sites in Central London. Secondly, the proposed modelling approach only utilises the NO_2 levels from neighbouring sites to estimate the NO_2 levels OF a given target site. Most of the aforementioned works utilise other explanatory variables that can mask the influence and effectiveness of their proposed technique on the overall model results.

The remainder of the paper is organised as follows: Section 2 presents the collected data and site locations. Section 3 describes the methods used and proposed modelling approach. Section 4 provides the details of the settings of the experiments carried out in this paper. Section 5 presents and discusses the results. Finally, Section 6 concludes the paper.

2. Data and Case Study Locations

2.1. Site Locations

The proposed models were built using hourly NO_2 concentration data collected at several locations in Central London. In more detail, the monitoring sites surrounding London Marylebone Road were chosen as the said road has breached the legal limits of NO_2 levels multiple times in recent years (King's College London, 2019). The number of sites considered in this study was limited to five, representing those closest to London Marylebone Road site. The following sites were selected: (1) London Marylebone Road (MAR), (2) North Kensington (KEN), (3) Camden Kerbside (CAM), (4) London Bloomsbury (BLM), (5) London

Westminster (WST), and (6) Tower Hamlets Roadside (HAM) (see Figure 1). The sites are part of the Automatic Urban and Rural Network (AURN) that has been monitoring concentration levels of ambient air pollutants such as ozone, oxides of nitrogen, carbon monoxide, sulphur dioxide and fine particles since 1997 (DEFRA, 2004).

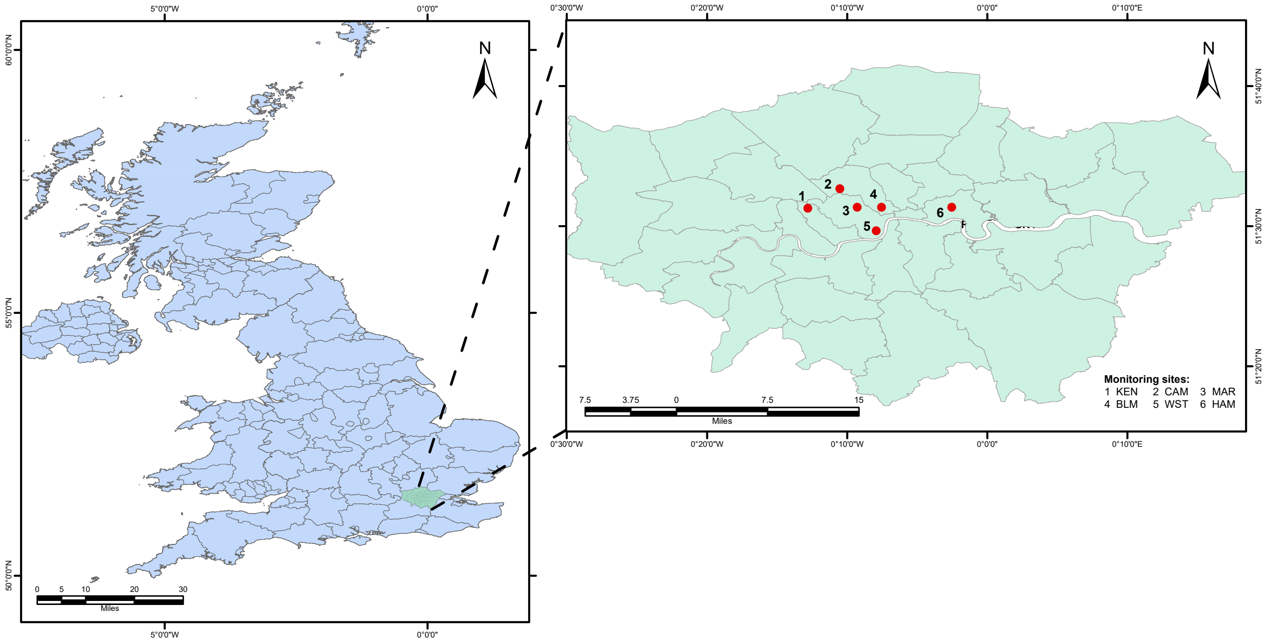


Figure 1: Area of study and selected air quality monitoring sites.

The sites of urban type are chosen as they normally exhibit a wide range of air pollution concentration values which is appropriate for testing the proposed ANN models (see Table 1). Furthermore, sites with missing values of less than or equal to five percent of the total number of data between 2013 and 2014 were selected.

Table 1: Air quality monitoring sites, coordinates, and site environment type.

Site name	Site Code	Latitude	Longitude	Altitude (m)	Distance from road (m)	Environment type
London Marylebone Road	MAR	51.522530	-0.154611	35	1	Urban traffic
North Kensington	KEN	51.521050	-0.213492	5	5	Urban background
Camden Kerbside	CAM	51.544210	-0.175269	50	3	Urban traffic
London Bloomsbury	BLM	51.522290	-0.125889	20	25	Urban background
London Westminster	WST	51.494670	-0.131931	5	17	Urban background
Tower Hamlets Roadside	HAM	51.522530	-0.042155	20	6	Urban traffic

Figure 2 shows the maps describing the selected sites and their vicinity. The key characteristics of the said sites are described as follows:

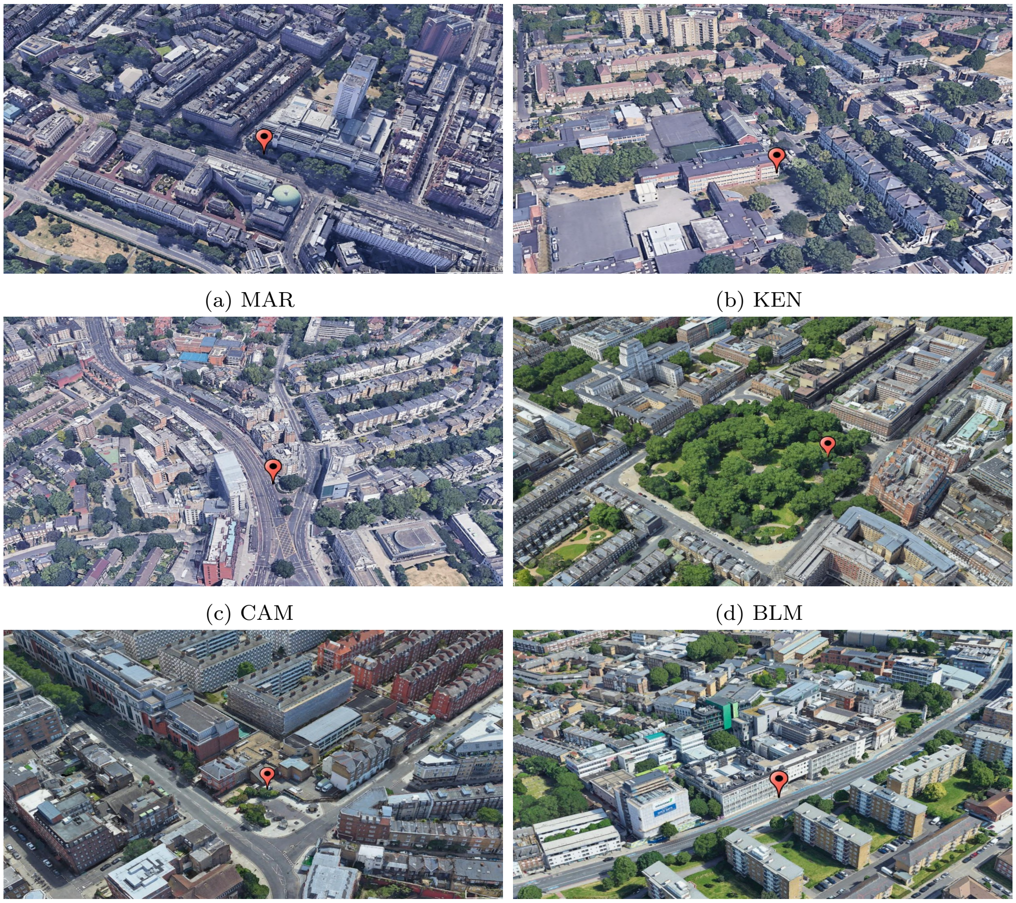
1) MAR is a kerbside site located within one meter of the edge of a busy six-lane road, A501. Its surrounding area forms a canyon;

2) KEN is a background site situated at a mainly residential area and is 5 meters away from a quiet residential road, St. Charles Square. CAM is a kerbside site situated at the southern end of a broad street canyon where the road is often busy;

3) BLM is a background site situated within the north-east corner of a central London garden with all four sides surrounded by a two-lane one-way road system. The site is surrounded by small buildings;

4) WST is a background site situated in the car park of a building, Westminster Coroner's Court, 17 meters away from an intersection between two-lane roads, B323 Horseferry Road and Regency Street. The site is surrounded by a mix of commercial and residential areas; and

5) HAM is a kerbside site situated within an existing building, part of Queen Mary and Westfield College, on a busy dual carriageway road, A11 Mile End Road. Its surrounding area consists of commercial and residential buildings.



(a) MAR

(b) KEN

(c) CAM

(d) BLM

(e) WST

(f) HAM

Figure 2: Maps describing the selected sites and their vicinity (via Google Maps).

2.2. Collected data

Hourly NO_2 concentration levels from January 2013 to December 2014 are obtained via an online resource run by AURN. All gathered data have been quality-assured or ratified, i.e. have undergone additional reviews so that faulty values are excluded (DEFRA, 2004). A total of 105,120 data points were collected in total, e.g. 17,520 data points from each chosen site. Chemiluminescent analysers were utilised in the continuous measurements of NO_2 concentrations.

Table 2 presents the measure of central tendency and dispersion of the collected NO_2 concentration data from all chosen sites. It is evident that MAR, CAM and HAM sites suffer from high NO_2 concentrations with recorded mean values of 89.18 μg/m^3, 68.10 μg/m^3 and 61.29 μg/m^3 during the period 2013-2014. In fact, breaches of the legal limit were observed 127 and 56 times at MAR and CAM sites, respectively, while a breach was recorded only twice at HAM site. On the other hand, the annual mean NO_2 concentration levels at KEN, BLM and WST sites were significantly lower, e.g. 35.63 μg/m^3, 51.46 μg/m^3, and 45.40 μg/m^3, respectively, compared to those from the roadside sites. Finally, the missing NO_2 concentration data in this study ranged from approximately 0.5% to 2.0%

Table 2: Descriptive statistics of the collected hourly NO_2 concentration data.

	Sites					
	MAR	KEN	CAM	BLM	WST	HAM
Mean [μg/m^3]	89.18	35.63	68.10	51.46	45.40	61.29
Median [μg/m^3]	82.99	31.03	62.46	50.27	43.62	59.36
Maximum [μg/m^3]	280.74	173.73	368.86	192.43	174.72	237.79
Minimum [μg/m^3]	7.25	0	5.56	1.14	0.29	0.92
Standard deviation [μg/m^3]	40.30	21.32	35.58	22.47	22.57	28.63
Missing data [%]	1.32	1.63	0.47	1.26	1.03	1.79

The observations above are consistent with the plots of the hourly variations and frequency distribution of the collected NO_2 concentration data shown in Figures 3 and 4, respectively.

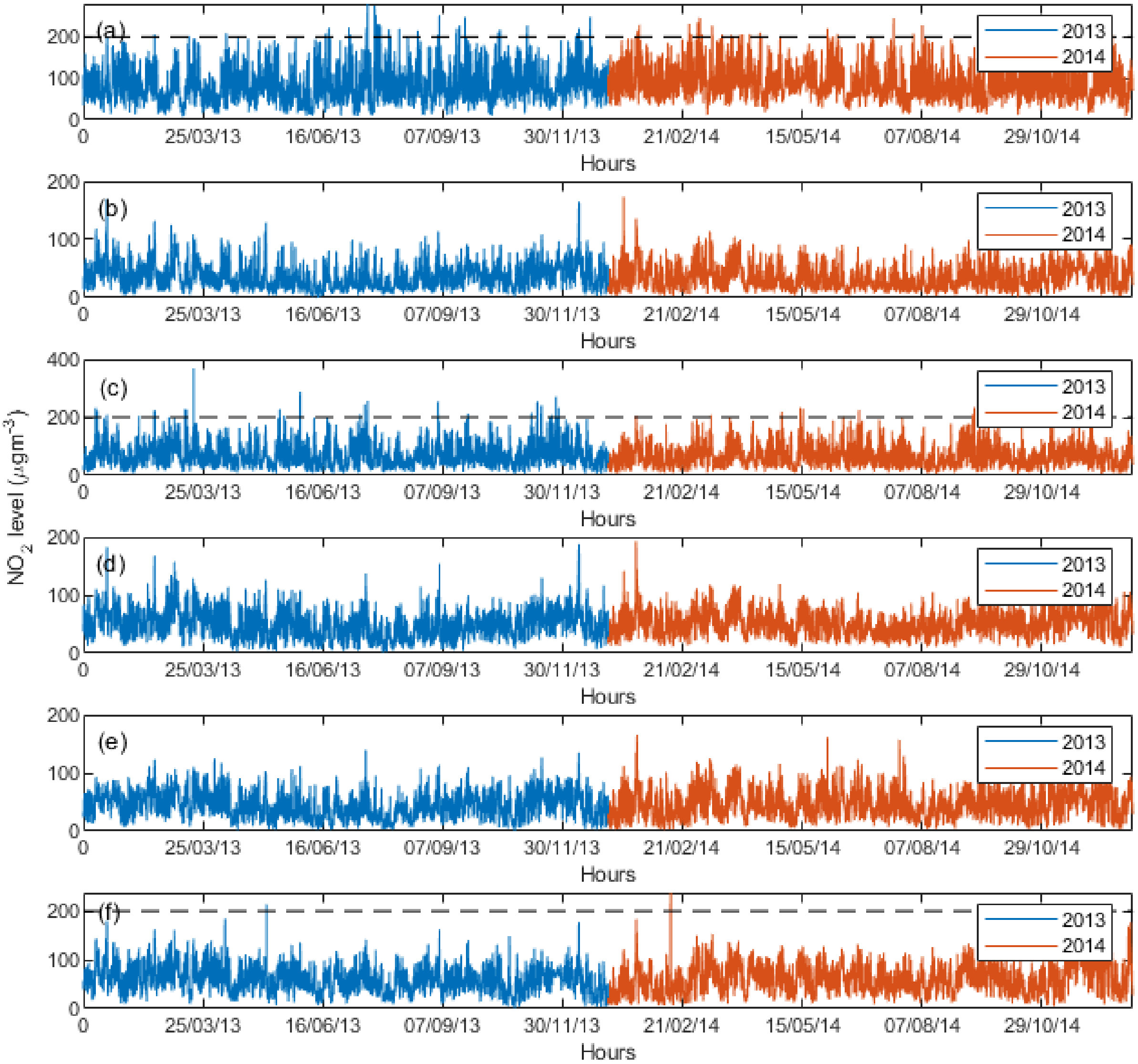


Figure 3: Hourly variations of the collected NO$_2$ concentrations at (a) MAR, (b) KEN, (c) CAM, (d) BLM, (e) WST, and (f) HAM sites from January 2013 to December 2014, where the dashed lines denote the limit value of 200 μg/m^3 set by the EU Air Quality Standards (DEFRA, 2004).

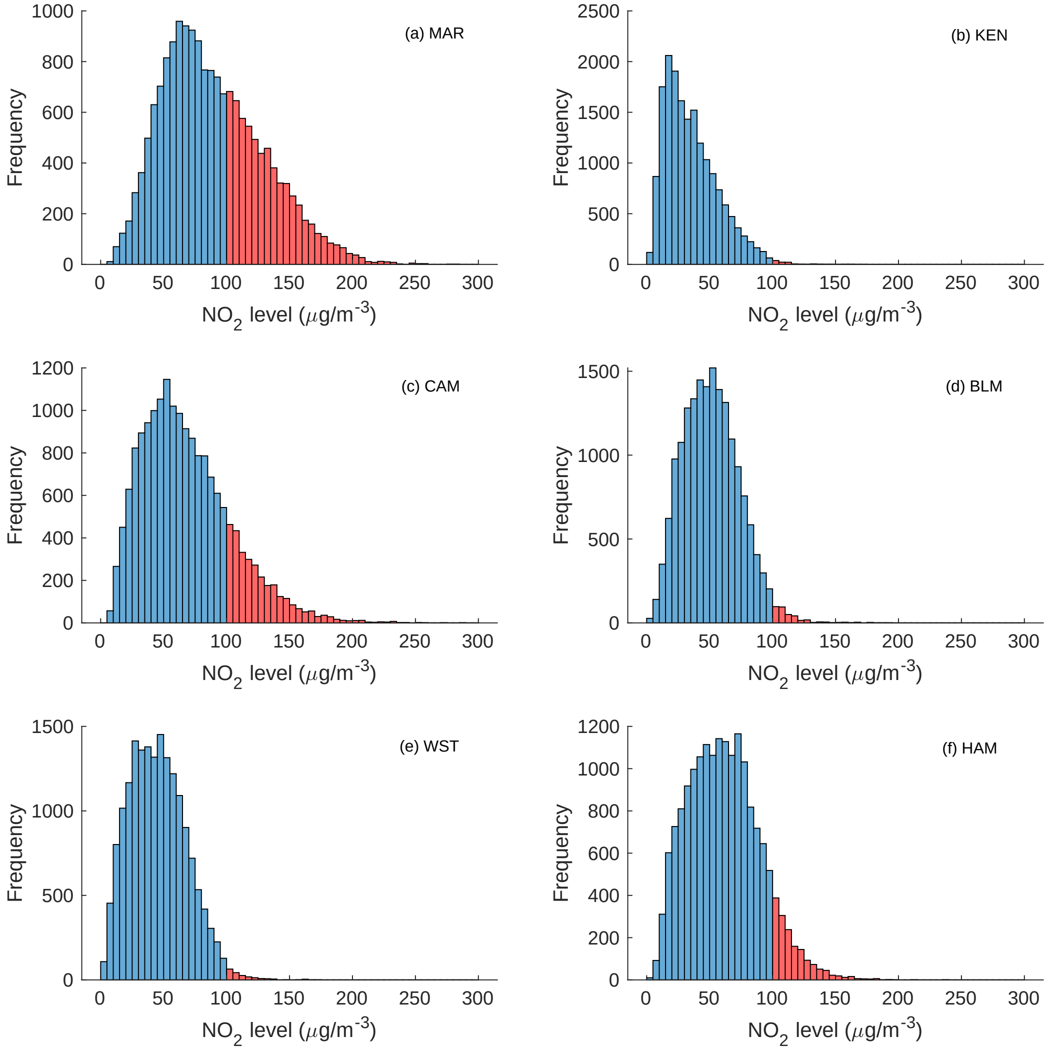


Figure 4: Frequency distributions of hourly NO_2 concentrations at (a) MAR, (b) KEN, (c) CAM, (d) BLM, (e) WST, and (f) HAM sites collected from January 2013 to December 2014 (DEFRA, 2004).

The plots of the mean hourly NO_2 concentration data during the study period are depicted in Figure 5. Consistent to the values presented in Table 1, the NO_2 concentration levels at the kerbside sites, e.g. MAR, CAM and HAM sites, are significantly higher than those from the background sites, e.g. KEN, BLM and WST. The NO_2 levels at all stations also demonstrate a consistent pattern, which is characterised by one peak in the morning, between 07:00 to 10:00, and another in the late afternoon, between 16:00 to 18:00, at the kerbside sites. The same trend can be observed between 19:00 to 21:00 at the background sites. On the other hand, stable low levels are observed at 04:00 at all sites. The observed trends demonstrate the strong influence of road transportation especially from diesel vehicles on roadside NO_2 levels (Colls, 2001; DEFRA, 2004; WHO, 2003).

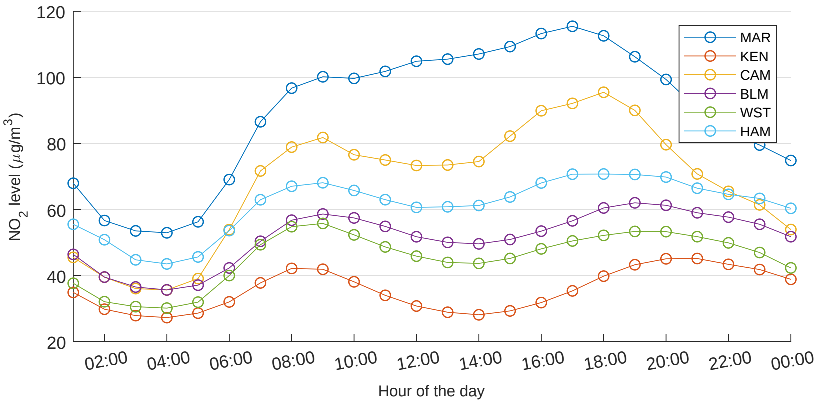


Figure 5: Mean hourly average values of the NO_2 levels at the six selected sites measured from January 2013 to December 2014 (DEFRA, 2004).

To further examine the variability of the collected data, the ratio of the standard deviation and mean value of collected time series, e.g. SD/Mean, and the signal-to-noise ratio (SNR), defined in decibels as $\text{SNR} = 20 \log(\text{Mean}/SD)$, were also calculated. As presented in Table 3, the SD/mean ratios of the data from KEN and CAM sites for both 2013 and 2014 are relatively higher than those calculated from other sites. This observation is also consistent with the computed SNR values from the data collected at KEN and CAM sites. These results indicate the potential difficulty in forecasting NO_2 levels using the said datasets as data variability directly contributes to the difficulty of a prediction task (Siwek and Osowski, 2012).

Table 3: Variability analysis of the collected NO_2 concentration data at the six selected sites.

	2013		2014	
Sites	SD/Mean	SNR (dB)	SD/Mean	SNR (dB)
MAR	0.47	6.47	0.42	7.50
KEN	0.60	4.49	0.58	4.62
CAM	0.54	5.41	0.50	6.00
BLM	0.47	6.52	0.39	8.07
WST	0.49	6.18	0.50	6.07
HAM	0.43	7.29	0.49	6.15

3. Methodology

3.1. Discrete Wavelet Transformation

Discrete wavelet transformation (DWT) is a popular signal processing technique that decomposes a given time series into several subseries of various scales (Mallat, 1989; Nievergelt, 2013). The initial step of DWT is to map the elements of a given time series S to its wavelet coefficients, and from these coefficients, two components are formed, namely a smooth version called approximation and a component corresponding to the deviations called details of the signal. The said process is described by the following expression:

$$S(t) = \sum_{k=1}^{n} c_{j,k}\psi_{j,k}(t) + \sum_{j=1}^{J}\sum_{k=1}^{n} d_{j,k}\psi_{j,k}(t), \tag{1}$$

where $\psi_{j,k}(t)$ and $\varphi_{j,k}(t)$ represent the mother wavelet and binary scale functions, respectively, $c_{j,k}$ and $d_{j,k}$ denote the approximation and detailed coefficients, respectively, at scale j and location k, n is the size of the original time series, and J is the decomposition level.

For instance, a decomposition of $S(t)$ into a low frequency part $A_1(t)$ and a high frequency part $D_1(t)$ is given by $S(t) = D_1(t) + A_1(t)$. As shown in Figure 6, the same process is carried out on $A_1(t)$ in order to obtain decomposition in finer scales: $A_1(t) = D_2(t) + A_2(t)$. Hence, Eq. (1) can be simplified into

$$S(t) = \sum_{i=1}^{J} D_i(t) + A_J(t), \tag{2}$$

where $D_i(t) = \sum_{k=1}^{n} d_{j,k}\psi_{j,k}(t)$ and $A_J(t) = \sum_{k=1}^{n} c_{j,k}$. In other words, DWT represents S in terms of the sum of subseries consisting of high frequency detail signals $D_1, D_2, \ldots, D_J$ and a low frequency approximation signal A_J (Mallat, 1989).

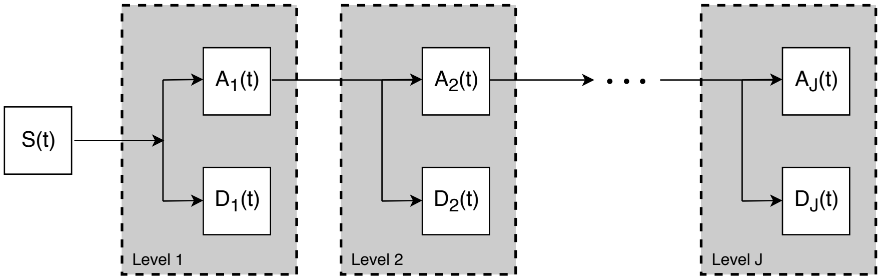


Figure 6: Decomposition of time series S into approximation A_J and detailed components D_i $i \in [1, J]$.

The performance of a wavelet transformation technique generally relies on the choice of the wavelet function, ψ, and the number of decomposition levels, J. In this study, the Daubechies wavelets (Daubechies, 1988) was chosen to approximate the raw time series signal for various reasons. Firstly, Daubechies wavelets are a family sophisticated wavelets capable of approximating continuous signals more accurately with fewer fixed scaling functions (Nievergelt, 2013). Secondly, Daubechies wavelets have been found to perform well in the past (Siwek and Osowski, 2012; Osowski and Garanty, 2007; Dunea et al., 2015). The names of the Daubechies family wavelets are usually written as DbN_v, where N_v is the number of vanishing moments which determine the ability a wavelet to approximate any given signal (Nievergelt, 2013).

A sample 5-level decomposition of the first one thousand hourly NO_2 observations at MAR site is shown in Figure 7. It can be observed that the higher is level of the wavelet, the lower is the variation of the detailed and approximation coefficients. As such, the wavelet decomposition technique can improve the performance of an ANN approximator in estimating the collected NO_2 concentration time series.

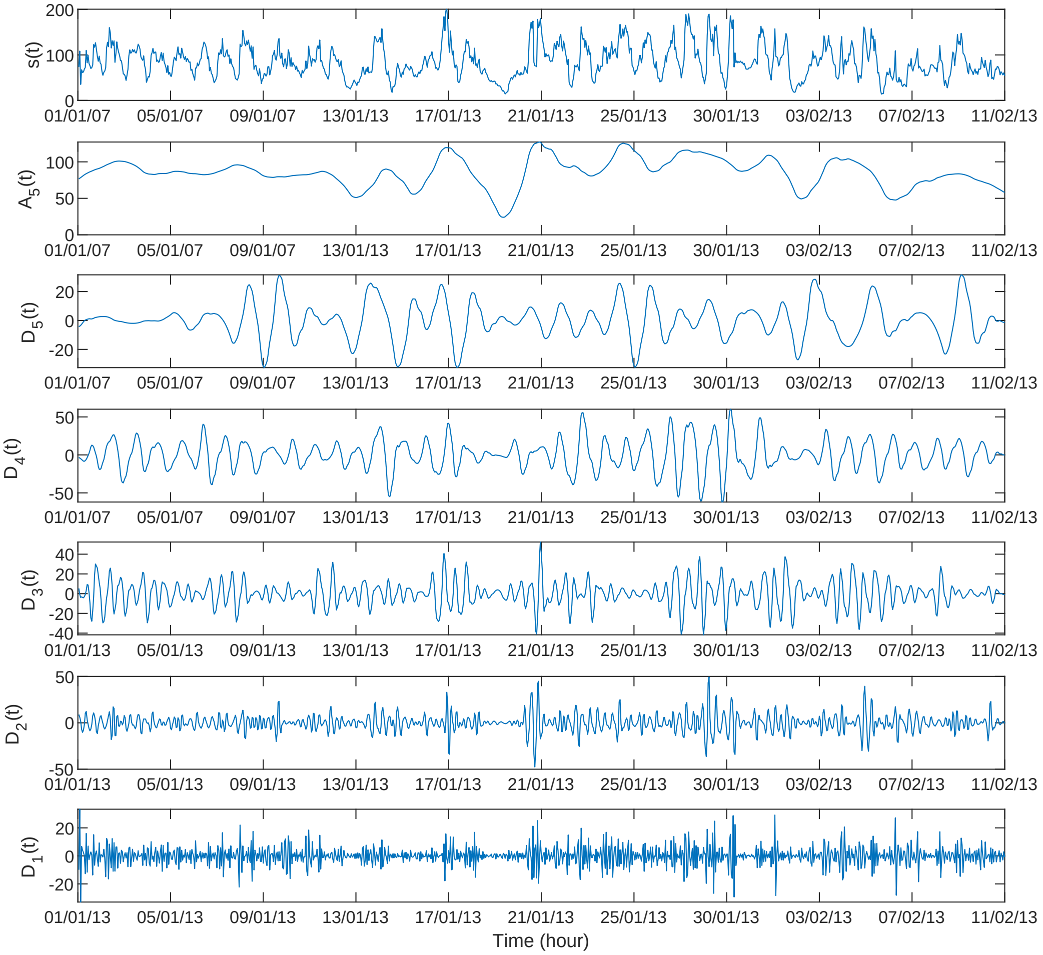


Figure 7: Wavelet decomposition of a subset of the collected NO$_2$ time series data, s(t); D$_1$ to D$_4$ represent the detailed coefficients, and A$_4$ the coarse approximation of s(t) on the fifth level.

3.2. Feed-forward ANNs

ANNs are computational structures inspired by the architecture and information-processing characteristics of a biological neural network (McCulloch and Pitts, 1943). ANNs perform a non-linear parametrised mapping F from an input $\mathbf{x}$ to an output $\mathbf{y}$,

$$\mathbf{y} = F\left(\mathbf{x}, \mathbf{w}\right), \tag{3}$$

where $\mathbf{w}$ denotes the weights and biases of the network. ANNs usually consist of single input and output layers, and one or more hidden layers, each of which has a varying number of interconnected neurons. A sample diagram of an ANN with N input neurons, and single hidden and output layers with M and K neurons, respectively, is shown in Figure 8.

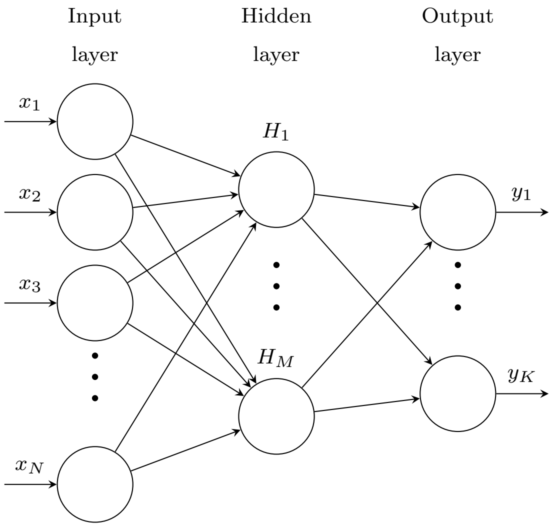


Figure 8: A feed-forward ANN with N input, M hidden, and K output neurons.

One popular form of ANNs in the field of air pollution modelling fall under the feed-forward type (Cabaneros et al., 2019; Gardner and Dorling, 1998; Shahraiyni and Sodoudi, 2016). In feed-forward ANNs, the information moves from the input to the succeeding layers in a single direction. Each node in the hidden layer is initially fed with the model predictors $x_1, \ldots, x_N$. Each input is scaled and shifted by the weights and bias parameters, respectively. The resulting value is then mapped by a non-linear function to the nodes of the succeeding layers. The main processes of a feed-forward ANN are mathematically described in Eq. (4), adapting the notation by Bishop (1995):

$$a_j = f\left(\sum_{i=1}^{N} x_i w_{ji}^{(1)} + b_j^{(1)}\right), \tag{4}$$

where $j \in [1, M]$, and M is the number of hidden nodes, $w_{ji}^{(1)}$ and $b_j^{(1)}$ are the weights and bias parameters, respectively, $f(\cdot)$ is a continuous real mapping, or the transfer function, and the superscript $^{(1)}$ denotes that the corresponding parameters are those in the first layer of the network. Following Eq. (4), the final result of the output layer of an ANN with only one hidden layer can be computed as follows:

$$y_k = \sum_{j=1}^{N} a_j w_{kj}^{(2)} + b_k^{(2)}, \tag{5}$$

where $k \in [1, K]$, and K is the number of output nodes.

ANNs are trained in a supervised manner where a series of input and desired output values are fed to the model. During the process, the weights and bias parameters are calibrated based on the network error, e.g. the difference between the input and target values. With the objective to minimise the overall network error, the training process is performed repeatedly according to a gradient descent algorithm until a stopping criterion is met. A more detailed discussion on ANNs and their training algorithms can be found in Bishop (1995) and Hagan et al. (1995).

3.3. Long short-term memory neural network

LSTM neural networks are sophisticated ANN models that utilise several hidden layers with nodes that are self-connected allowing a cyclic flow of information (Hochreiter and Schmidhuber, 1997). An LSTM network contains one input layer, one output layer, and a series of memory blocks. Each memory block is composed of one or more self-recurrent memory cells and three multiplicative units, i.e. input, output and forget gates, that provide continuous analogs of read, write and reset operations for the blocks. The blocks enable the network to preserve enough information to update their training parameters. The said components enable LSTM networks to overcome the vanishing gradient problem in which a network stops learning from previous temporal patterns of a given data due to multiple gradient updates. As such, LSTM networks are suitable for time series forecasting applications (Freeman et al., 2018; Hochreiter and Schmidhuber, 1997).

An LSTM memory block with a single cell is illustrated in Figure 9.

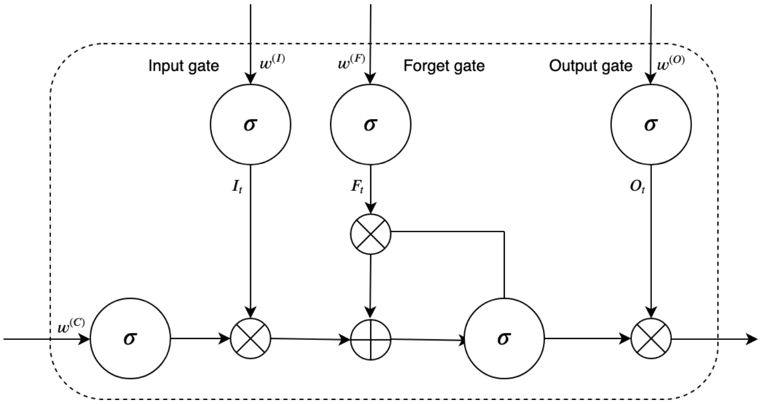


Figure 9: A diagram of a single LSTM memory block

The input gate allows incoming information to modify the state of the nodes. The output gate permits or impedes the cell state from affecting other neurons. The forget gates were designed to learn and reset memory cells once their status is out of date, thereby preventing the cell status from growing without bounds and causing saturation of the transfer functions. The forward training process of an LSTM unit can be formulated as described in Eq. (6) to (10):

$$F_t = \sigma\left(w^{(F)}\sum_{k=1}^{N} h_{t-1}\cdot x_{k,t} + b^{(F)}\right) \tag{6}$$

$$I_t = \sigma\left(w^{(I)}\sum_{k=1}^{N} h_{t-1}\cdot x_{k,t} + b^{(I)}\right) \tag{7}$$

$$c_t = F_t\cdot c_{t-1} + I_t\cdot\sigma\left(w^{(C)}\sum_{k=1}^{N} h_{t-1}\cdot x_{k,t} + b^{(C)}\right) \tag{8}$$

$$O_t = \sigma\left(w^{(O)}\sum_{k=1}^{N} h_{t-1}\cdot x_{k,t} + b^{(O)}\right) \tag{9}$$

$$h_t = O_t \cdot \sigma\left(c_t\right) \tag{10}$$

where I_t, O_t, and F_t are the outputs of the input, output and forget gates at time t, respectively, c_t and h_t represent the activation vector for each cell and memory block, respectively, σ denotes the transfer function, and w and b are the weighting and bias constants.

3.4. Framework of the proposed modelling approach

The overall schematic of the proposed modelling approach is presented in Figure 10. The proposed modelling approach consists of four main components, each of which is briefly described as follows:

(1) The approach implements the leave-one-out cross-validation methodology in which a particular site is chosen as the target site and only the data from the remaining neighbouring sites are utilised to spatially estimate the concentration levels of the target site.

(2) The DWT technique applied to decompose the raw NO_2 concentration time series into several $(J+1)$ sub-series as described by Eq. (2). The main idea of the approach is to develop several models to estimate the said sub-series with lower variability instead of employing a single model to estimate the original time series exhibiting higher variability.

(3) LSTM models are developed to extract the long-term temporal characteristics from each of the $(J+1)$ decomposed sub-series.

(4) Given a total number of k monitoring sites, the individual LSTM model estimates are then reconstructed using the following expression to calculate the final forecasting results:

$$\hat{Y}\left(t\right) = F_1(D_1(t-h)) + F_2(D_2(t-h)) + \ldots + F_J(D_J(t-h)) + F_{J+1}(A_J(t-h)), \tag{11}$$

where $\hat{Y}\left(t\right)$ is the estimated pollutant concentration at time t, F is the estimator represented by the LSTM model, D_j and A_J, for $j = 1, 2, \ldots, J$, represent the detailed and approximation coefficients of the pollutant concentration from the $(k-1)$ neighbouring sites, respectively, and h is the forecasting horizon.

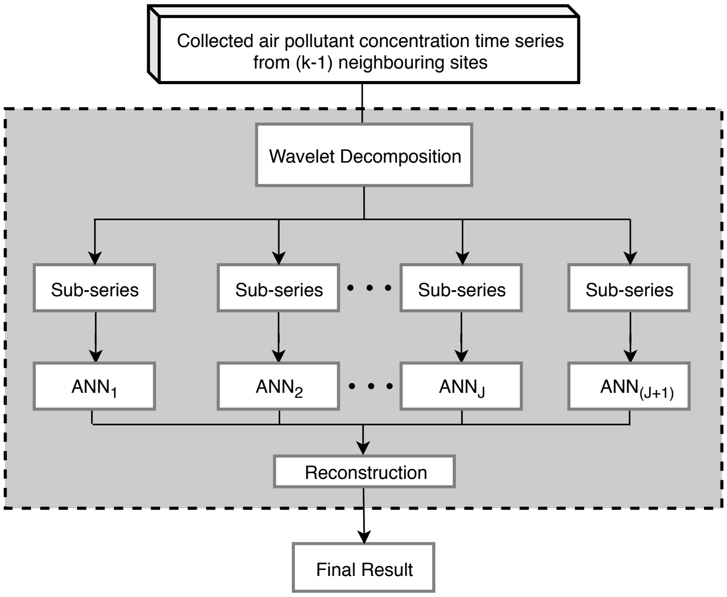


Figure 10: The framework of the proposed spatio-temporal forecasting model.

4. Model Development

Building ANN models generally has seven main processes, namely: (1) selection of predictors, (2) pre-processing of data, (3) splitting of data, (4) selection of model architecture, (5) optimisation of model structure, (6) calibration of model parameters, and (7) evaluation of model results (Cabaneros et al., 2019). The implementation of the said processes in this study is briefly described as follows.

4.1. Predictor selection

Hourly NO_2 concentration data collected from the $(k-1)$ neighbouring sites are utilised to estimate the NO_2 concentration data of a given target site. Furthermore, the significant neighbouring sites are determined by calculating the correlation between the data from the target and neighbouring sites. Finally, the number of predictor lags for each reference site was then determined after computing the effect of various lags on the autocorrelation function values of the collected NO_2 levels from each site.

4.2. Data pre-processing

The problem of missing data was addressed using the Data Augmentation (DA) algorithm (Tanner and Wong, 1987). DA algorithm alternatively replaces the missing data with randomly assumed values of the parameters. It then makes inferences about the unknown parameters from a Bayesian posterior distribution based on the observed and imputed data. A more detailed discussion about the DA algorithm can be found in Tanner and Wong (1987). It is also a common practice to normalise the collected raw data to be normalised before utilising them for model development. As such, the data were max-min normalised to ensure that all values fall between zero to one (Hagan et al., 1995).

4.3. Data splitting

The available dataset from each site is then split into three subsets, namely, the training, validation and testing subsets to avoid model overfitting. The hourly data measured from 1 January 2013 to 26 May 2014 (12,264 points) were utilised for the calibration of the network weights and bias parameters, while those measured from 27 May 2014 to 31 December 2014 (5,256 points) were used for both network validation and testing purposes. It is worth noting that the number of samples depends on the prediction horizon, h. That is, the farther out the prediction, the fewer samples will become available due to the effects of time-shifting. Hence, the total amount of remaining samples is equal to 17,520–h. In this study, the hourly NO_2 concentration was estimated 1 hour ahead.

4.4. Model architecture selection

As described in Section 3.4, a data-intensive hybrid model based on DWT and a deep learning LSTM model was developed in this study. A popular feed-forward ANN model, the Multilayer Perceptron (MLP), was also developed to serve as a benchmark of the proposed hybrid model. Furthermore, the logistic sigmoid and linear functions were used as activation functions in the hidden and output layers, respectively. The former was selected because it is nonlinear and continuous, attributes that enable ANNs to be capable of approximating any smooth and measurable function (Hornik et al., 1989). On the other hand, the latter was selected because it yields continuous and unbounded values, attributes that are considered as appropriate for approximation and regression tasks (Hagan et al., 1995).

4.5. Model structure optimisation

One hidden layer was employed in the proposed model as it is found to be sufficient in approximating any smooth measurable mapping between the predictor and target variables (Hornik et al., 1989). Additionally, the optimal number of nodes in the hidden layer was determined by a trial-and-error method. In detail, several models with various hidden layer configurations were run 100 times. This is carried out to account for the sensitivity of the training to the initial values of synaptic weights and biases. The configuration that yielded the least average model error, e.g. in terms of the mean absolute error (MAE), was considered optimal and selected for further model testing.

4.6. Model training

The model weights and bias parameters of the LSTM models are calibrated using the Adam algorithm (Kingma and Ba, 2014). On the other hand, the Levenberg-Marquardt (LM) algorithm (Levenberg, 1944; Marquardt, 1963) was utilised to train the MLP models. For a more comprehensive discussion on said algorithms, the reader is advised to see Bishop (1995) and Kingma and Ba (2014). In this paper, the training of the models with random initial weights and bias factors was repeated (100 repetitions) to account for the sensitivity of the algorithms to initial synaptic weights.

Table 4: Model performance metrics.

Abbreviation	Definition	Equation
RMSE	Root mean squared error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - O_i\right)^2}$
MAE	Mean absolute error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\left\|P_i - O_i\right\|$
R^2	Correlation of dependence	$\mathrm{R}^2 = \frac{\sum_{i=1}^{N}\left(P_i - \hat{O}\right)^2}{\sum_{i=1}^{N}\left(O_i - \hat{O}\right)^2}$
FB	Fractional bias	$\mathrm{FB} = 2\left(\frac{\bar{O}-\bar{P}}{\bar{O}+\bar{P}}\right)$

4.7. Performance evaluation

This study utilises the statistical indices listed in Table 4 for evaluating the performance of the models. Note that P_i and O_i are the predicted and observed values of NO_2 concentration, respectively, and $\bar{P}$ and $\bar{O}$ are the mean value of the predicted and observed values of NO_2 concentration, respectively.

The quality of the forecasting results is further assessed by the contingency table shown in Table 5. The columns are the forecast values and the rows are the actual values. In the matrix, TN is the number of non-episode days correctly identified, FP is the number of non-episode days incorrectly identified as an episode, e.g. false alarm, FN is the number of episode days incorrectly identified as non-episode days, and TP is the number of episode days correctly identified. The NO_2 level of 100 μg/m^3 was selected as it is found that levels beyond it are considered unhealthy for sensitive groups, including those with lung disease, children and older adults (USEPA, 2016).

Table 5: Contingency table for a two-category forecast.

		Forecast	
		< 100 μg/m^3	≥ 100 μ g/m^3
Actual	< 100 μg/m^3	TN	FP
	≥ 100 μg/m^3	FN	TP

Based on Table 5, several metrics can be calculated: the probability of detection (POD) and false alarm rate (FAR). POD represents the fraction of correctly forecast NO_2 episode days, ranging between 0 to 1 and with the best value of 1. FAR is the fraction of false alarms over the total forecast positive events, ranging between 0 to 1 and with the best value of 0. Ideally, the POD score should be reasonably high while the FAR score should be reasonably low to maintain public confidence in the NO_2 level early warnings.The said ratios are given by Eqs. (12),and (13), respectively:

$$POD = \frac{TP}{FN + TP}, \tag{12}$$

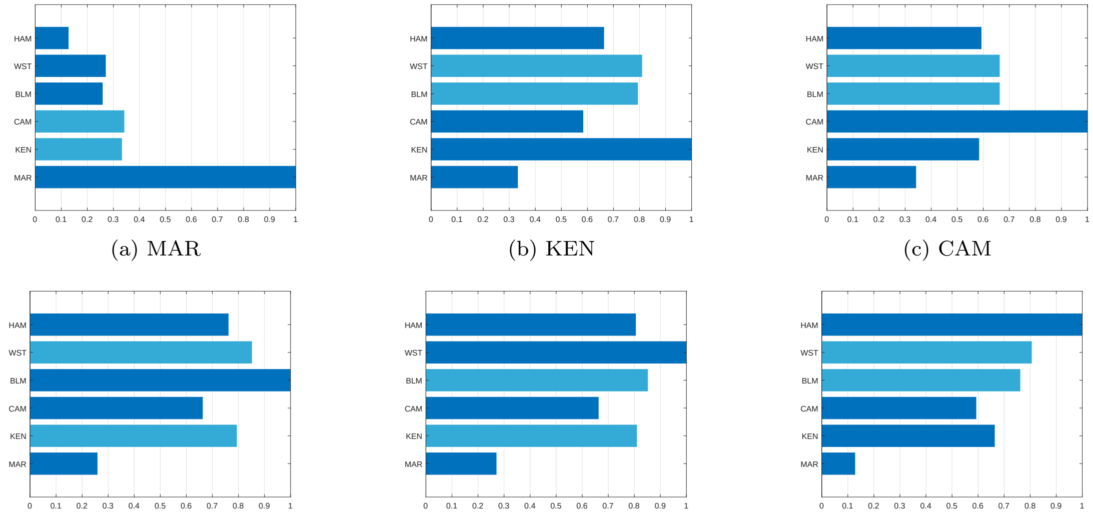
$$FAR = \frac{FP}{FP + TP}. \tag{13}$$

All computations described in this study are written and implemented in MATLAB R2018a software (The MathWorks, 2019).

5. Results and Discussion

5.1. Correlation and lag analysis

As shown in Figure 11, there is a high mutual correlation ($> 60\%$) between the NO_2 data collected from all sites but MAR site. The data from BLM and WST sites exhibit the highest mutual correlation index (85.2%). With the exception of MAR site, CAM and KEN sites obtained the lowest mutual correlation index (58.4%) despite their close proximity, e.g. approximately 2.55 km apart. It also appears that the data exhibiting the highest variability levels, e.g. data taken from MAR site, is least correlated with data collected from the remaining sites. Conversely, the dataset with the least variability level, e.g. data taken from KEN site, is highly correlated with the rest of the collected data.



(a) MAR (b) KEN (c) CAM

(d) BLM (e) WST (f) HAM

Figure 11: The correlation scores between the collected NO_2 concentration data from between a target site and neighbouring sites.

Ideally, the dataset that is least correlated with the rest of the datasets should be discarded. However, one of the objectives of the study is to test the spatial estimation ability of the proposed model at locations with a limited or poor quality of data. As such, variants of the wavelet-based and plain LSTM and MLP models are also built based on the predictors. Given the results provided in Figure 11, the top two neighbouring sites with data that are highly correlated to a given target site are chosen. Several benchmarks of the proposed model such as the plain and wavelet-based feed-forward ANN-models were developed. In this paper, a Multilayer Perceptron (MLP) model which is popular in the context of air pollution forecasting was chosen to represent the feed-forward benchmark model (Cabaneros et al., 2019; Gardner and Dorling, 1998; Shahraiyni and Sodoudi, 2016). Table 6 lists all models that were developed in this paper (note that CA denotes correlation analysis).

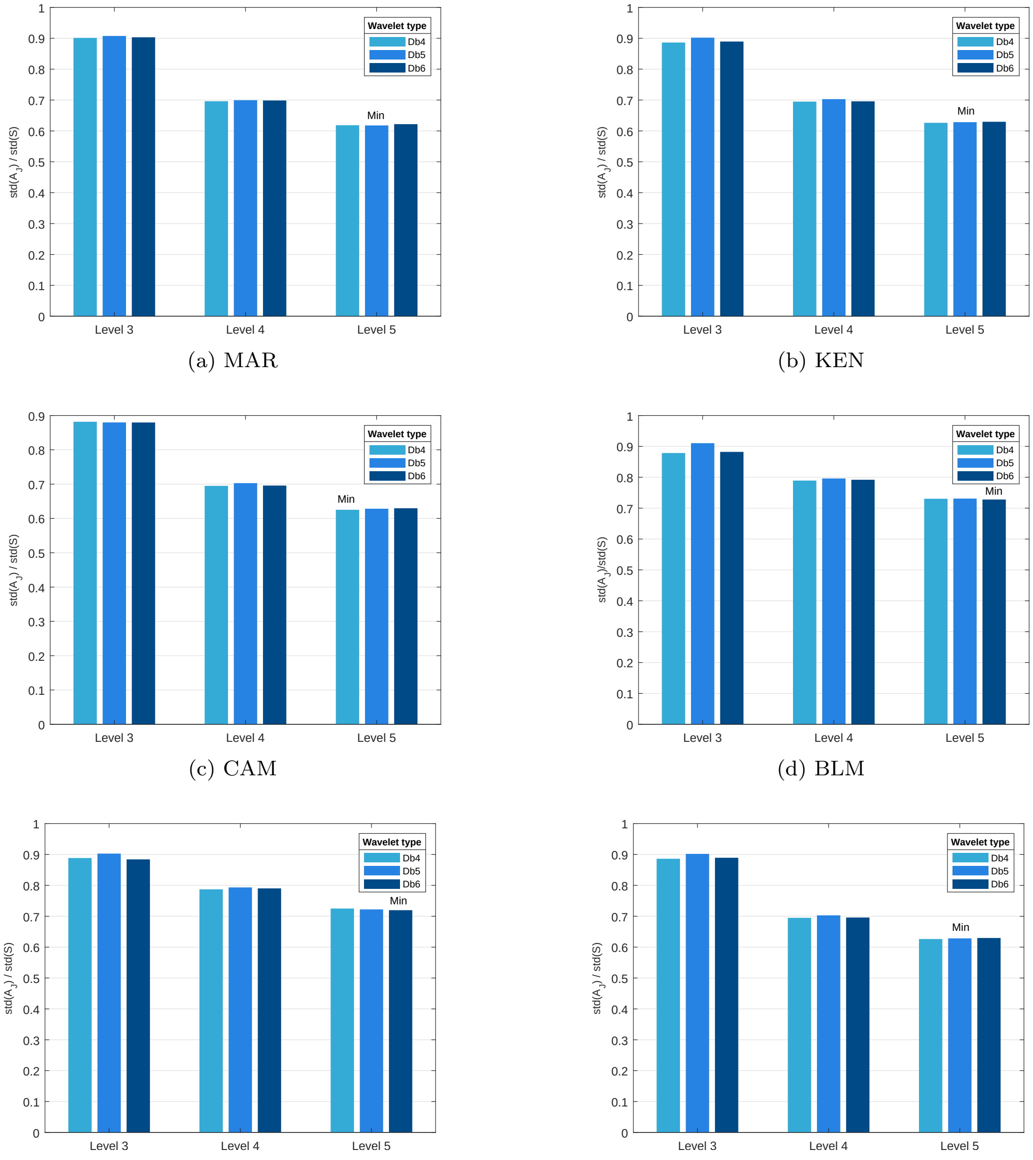
Table 6: Specification of the proposed and benchmark models.

Model code	Model description
W-LSTM-CA	Wavelet-based LSTM with predictors selected via CA
W-LSTM	Wavelet-based LSTM with all predictors
LSTM-CA	Plain LSTM with predictors selected via CA
LSTM	Plain LSTM with all predictors
W-MLP-CA	Wavelet-based MLP with predictors selected via CA
W-MLP	Wavelet-based MLP with all predictors
MLP-CA	Plain MLP with predictors selected via CA
MLP	Plain MLP with all predictors

For the lag analysis of the model predictors, the number of lags for each predictor was determined using the autocorrelation function. The computed optimum number of lagged inputs varies from 1 to 2. That is, for a specific target site j, the predictors utilised are $x_i\left(t-h\right)$, $x_i\left(t-h-1\right)$, and $x_i\left(t-h-2\right)$, where $i \in [1, 6]$ and $i \neq j$.

5.2. Daubechies wavelet selection

In the case of the wavelet-based models, several types and levels of the Daubechies wavelet (Daubechies, 1988) were tested and selected based on the ratio $\text{std}\left(A_j\right)/\text{std}\left(S\right)$. That is, the standard deviation of A_j must be substantially smaller than that of the original time series S. However, choosing a larger value of J increases the number of terms in Eq. (2), thus accumulating more approximation errors when Eq. (11) is computed (Osowski and Garanty, 2007). As such, the levels of the tested Db wavelets were limited from 3 to 5. On the other hand, the number of vanishing moments, N_v, was limited from 4 to 6. Considering the above-mentioned conditions, various configurations of Daubechies wavelets were selected. The results are shown in Figure 12. For models with target sites MAR through HAM, 5-level Db5, Db4, Db4, Db6, Db6, and Db5 wavelets provided the least $\text{std}\left(A_j\right)/\text{std}\left(S\right)$ ratios, respectively. As such, the following wavelet configurations were used. This also means that six ANN models should be trained: five for the detailed coefficients, Di, $(j = 1, 2, \ldots, 5)$ and one for the residual signal, A_5.



(a) MAR

(b) KEN

(c) CAM

(d) BLM

(e) WST

(f) HAM

Figure 12: Daubechies type and level optimisation results.

5.3. Model structure optimisation

Finally, the hidden layer configurations determined using the procedure described in Section 4 are summarised in Table 7. The optimal number of hidden nodes for the proposed LSTM models ranges from 100 to 150. More nodes tend to be needed to estimate the pollutant values at the kerbside target sites, e.g. MAR, CAM and HAM sites. Furthermore, a lesser number of hidden nodes is needed as the decomposition level using Daubechies wavelets increases, indicating that the variability of a given time series directly influences the complexity of the model needed to estimate the time series. On the other hand, the computed number of hidden nodes for the benchmark MLP models varies from 20 to 35 for the standalone MLP models and 20 to 50 for the wavelet-based MLP models. Similar to the results for the LSTM models, more hidden nodes are required to estimate the data collected from the kerbside urban sites. Lastly, the number of hidden nodes in most cases is less in models that use less number of predictors.

Table 7: Optimal hidden layer configurations of the both plain and wavelet-based MLP models.

Target	Predictors	Plain		W-MLP						W-LSTM					
Site		MLP	LSTM	D_1	D_2	D_3	D_4	D_5	A_5	D_1	D_2	D_3	D_4	D_5	A_5
MAR	All	35	150	20	30	20	25	20	30	125	115	115	110	115	110
	CAM; HAM	30	125	20	25	25	20	25	25	115	120	115	115	105	105
KEN	All	35	130	20	20	40	20	20	40	120	115	120	110	105	100
	BLM; WST	30	115	25	20	25	25	30	30	115	115	110	115	110	100
CAM	All	25	135	25	20	30	20	20	40	120	125	110	115	115	105
	BLM; WST	25	120	20	25	25	20	25	30	115	120	115	110	110	105
BLM	All	25	130	20	35	20	45	20	50	125	120	120	115	110	110
	KEN; WST	25	125	20	25	25	30	30	25	120	115	115	110	115	105
WST	All	30	130	25	20	25	20	25	40	125	125	115	120	115	115
	KEN; BLM	25	125	30	20	20	25	20	25	125	120	120	110	110	100
HAM	All	25	145	25	25	20	35	25	40	130	125	125	120	125	115
	BLM; WST	25	135	20	20	25	30	20	35	125	125	120	115	120	120

5.4. Overall model results

The results of the models on the testing data to forecast the NO_2 levels at MAR, KEN, CAM, BLM, WST, and HAM sites are presented in Table 8.

Table 8: The results of different models for 1h-ahead forecasting of hourly NO_2 levels.

		Target sites					
Model Code	Error Index	MAR	KEN	CAM	BLM	WST	HAM
W-LSTM-CA	RMSE $[\mu g/m^3]$	34.791	10.204	23.012	9.110	11.311	22.052
	MAE $[\mu g/m^3]$	25.021	7.241	18.001	6.177	8.006	16.122
	R^2	0.560	0.851	0.712	0.905	0.861	0.789
	FB	0.060	0.043	-0.127	0.049	-0.023	0.050
W-LSTM	RMSE $[\mu g/m^3]$	33.665	8.414	22.022	7.656	10.373	21.723
	MAE $[\mu g/m^3]$	24.500	5.808	16.785	4.892	7.086	15.224
	R^2	0.580	0.924	0.770	0.931	0.896	0.812
	FB	0.049	0.032	-0.113	0.036	-0.007	0.041
LSTM-CA	RMSE $[\mu g/m^3]$	34.911	10.831	23.621	10.135	12.051	22.178
	MAE $[\mu g/m^3]$	26.331	8.333	18.731	6.819	9.563	17.019
	R^2	0.556	0.851	0.711	0.899	0.859	0.781
	FB	0.062	0.041	-0.134	0.054	-0.031	0.057
LSTM	RMSE $[\mu g/m^3]$	34.224	9.610	22.708	9.579	10.854	22.730
	MAE $[\mu g/m^3]$	24.996	7.510	17.301	7.233	7.805	16.670
	R^2	0.572	0.891	0.749	0.893	0.881	0.797
	FB	0.051	0.045	-0.129	0.056	-0.013	0.064
W-MLP-CA	RMSE $[\mu g/m^3]$	34.899	11.019	23.981	9.953	12.377	23.423
	MAE $[\mu g/m^3]$	26.872	8.627	18.763	7.391	9.599	18.053
	R^2	0.557	0.847	0.708	0.884	0.843	0.778
	FB	0.071	0.036	-0.129	0.053	-0.012	0.069
W-MLP	RMSE $[\mu g/m^3]$	33.224	9.320	22.421	9.232	11.117	23.199
	MAE $[\mu g/m^3]$	24.796	7.324	17.116	7.115	8.371	18.031
	R^2	0.572	0.890	0.756	0.906	0.874	0.790
	FB	0.074	0.040	-0.117	0.050	-0.013	0.077
MLP-CA	RMSE $[\mu g/m^3]$	35.665	11.065	24.213	10.244	12.689	23.823
	MAE $[\mu g/m^3]$	27.029	8.680	19.110	7.673	10.003	18.432
	R^2	0.556	0.845	0.708	0.882	0.843	0.770
	FB	0.071	0.039	-0.138	0.063	-0.033	0.061
MLP	RMSE $[\mu g/m^3]$	36.654	9.664	22.924	10.186	11.201	23.695
	MAE $[\mu g/m^3]$	26.224	7.520	17.828	7.883	8.430	18.962
	R^2	0.564	0.880	0.741	0.881	0.872	0.783
	FB	0.056	0.057	-0.134	0.068	-0.010	0.068

It is seen that the integration of a wavelet decomposition technique and a deep learning model provides results superior to those of the benchmark models. The best model results are exhibited by the W-LSTM models for KEN and BLM sites. Both models yield the least RMSE and MAE values and can account for 92% to 93% of the variability of the actual NO_2 levels of the said sites. The W-LSTM models for CAM, WST and HAM sites also provide relatively satisfactory results, with R^2 scores ranging from 77% to 90%. In contrast, the MLP-CA model for MAR site provides the worst forecasting results, with R^2 score of only 58%. Furthermore, the deep learning models for all sites outperform the benchmark MLP models. For instance, a significant increase in forecasting accuracy ranging from 0.3% to 3.4% (R^2) is achieved by both the plain and wavelet-based LSTM models when compared to the MLP models. The W-LSTM models exhibited the greatest improvement in performance, e.g. 0.8% to 3.5% (R^2), when compared to the LSTM models. On the other hand, LSTM-CA models yielded the least accuracy improvement, e.g. 0% to 1.7% (R^2), when compared to the MLP-CA models.

5.5. Wavelet-based versus plain models

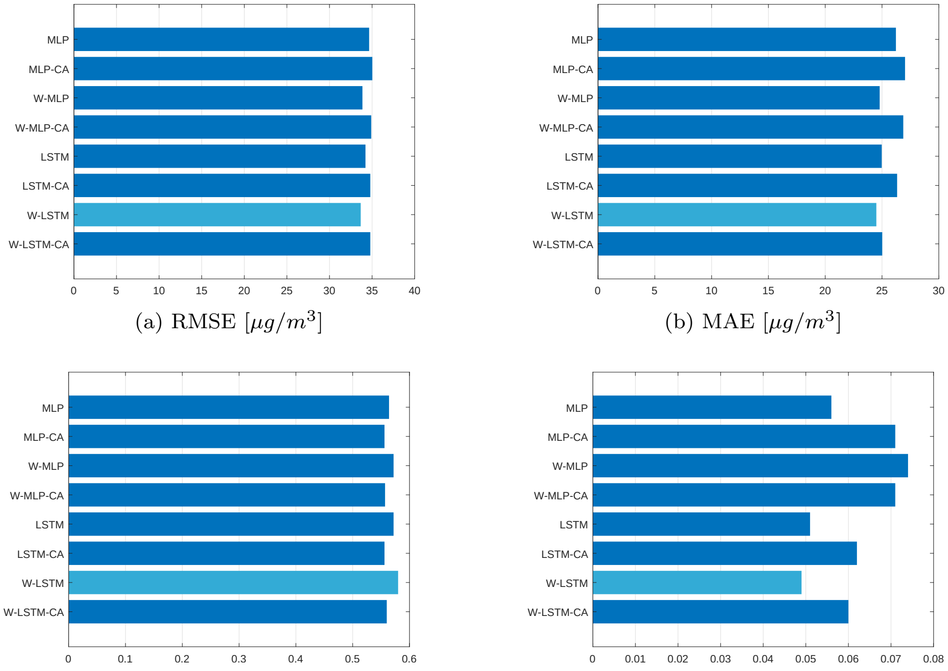
The results of the proposed and benchmark models for every target site are illustrated in Figures 13 to 18, where the highlighted bars represent the best-performing model. The overall performance of the wavelet-based models compares favourably with the plain models. For instance, an increase of 0.8% to 3.8% in R^2 score is achieved by applying DWT on the LSTM models, and 0.2% to 2.5% on MLP models. It worth noting that the use of wavelets on the LSTM models for MAR and WST sites only offered slight improvements on forecasting accuracy. A similar observation can be made for the MLP models for MAR, WST and HAM sites. Furthermore, less significant improvement is achieved by applying DWT on models utilising the predictors selected through correlation analysis. For instance, both LSTM and MLP models with two predictors only achieved a maximum improvement of 0.6% to 0.8% in terms of R^2 scores.

5.6. All predictors versus selection using correlation analysis

The reduction of some utilised predictors through correlation analysis tends to degrade the overall performance of the models. For instance, the greatest decrease in accuracy, e.g. 7.3% in R^2, was observed by the W-LSTM model for KEN site when only two predictors are utilised. A similar observation can be made for the remaining models where the decrease in R^2 scores ranges from 1.2% to 5.8%. The results above suggest that the information derived from all selected reference sites is important in helping the ANN models approximate the NO_2 levels at a given target site. However, it is worth noting that the models with a reduced number of predictors for KEN, BLM and WST sites still provided satisfactory results, with results explaining 85% to 90% of the variance of the actual NO_2 data. This indicates the applicability of the proposed spatiotemporal model in cases where the number of neighbouring sites is severely limited.

5.7. Site-dependency of model the results

It is also very apparent that the performance of the models is site-dependent, consistent with the findings of several previous works (Alimissis et al., 2018; Tzanis et al., 2019). Several factors such as the traffic characteristics, location, pollution sources, and geometry of the buildings around the target site tend to explain the results above. The models for the background sites, e.g. KEN, BLM and WST sites, significantly outperform the models for the kerbside sites, e.g. MAR, CAM and HAM sites. For instance, both plain and wavelet-based models for KEN, BLM and WST sites obtained the lesser error scores than those for MAR, CAM and HAM sites. It is also worth the emphasis that the performance of the plain LSTM and MLP models for KEN and BLM sites are almost similar. In summary, the ranking of the sites in terms of the model performance (in decreasing order) is as follows: BLM, KEN, WST, HAM, CAM and MAR (see Table 8). It can be said that the overall model results are influenced by the level of variability of the data. That is, the models for the target sites with dataset exhibiting high variability perform poorly. For instance, the ranking of the sites in terms of standard deviation values (in increasing order) almost matches the ranking above: KEN, BLM, WST, HAM, CAM and MAR (see Table 1). Furthermore, the mutual relationship between the data from the neighbouring and targets sites has a significant effect on the model results. That is, the NO_2 level data from KEN, BLM and WST sites are highly correlated, e.g. correlation score from 0.58 to 0.85, with the data of the remaining sites except MAR (see Figure 11).



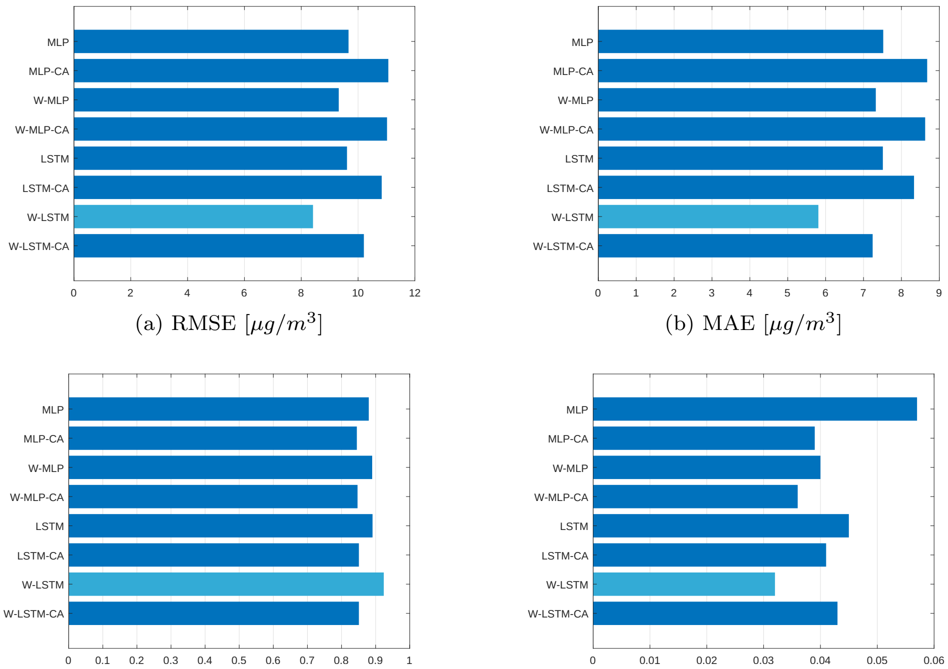
(a) RMSE [$\mu g/m^3$]

(b) MAE [$\mu g/m^3$]

(c) R^2

(d) FB

Figure 13: Forecasting results for MAR site.



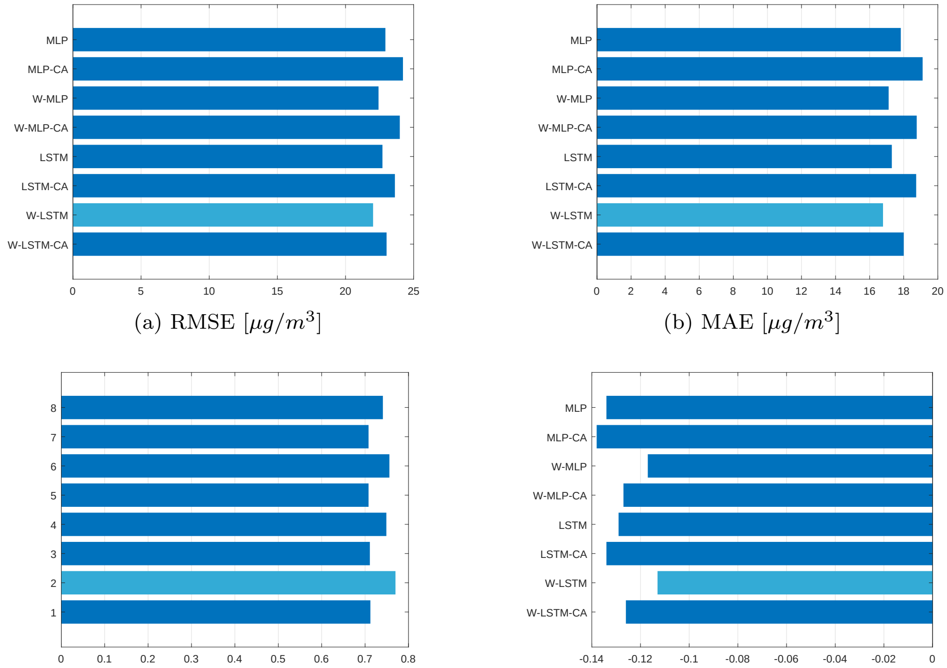
(a) RMSE $[\mu g/m^3]$

(b) MAE $[\mu g/m^3]$

(c) R^2

(d) FB

Figure 14: Forecasting results for KEN site.



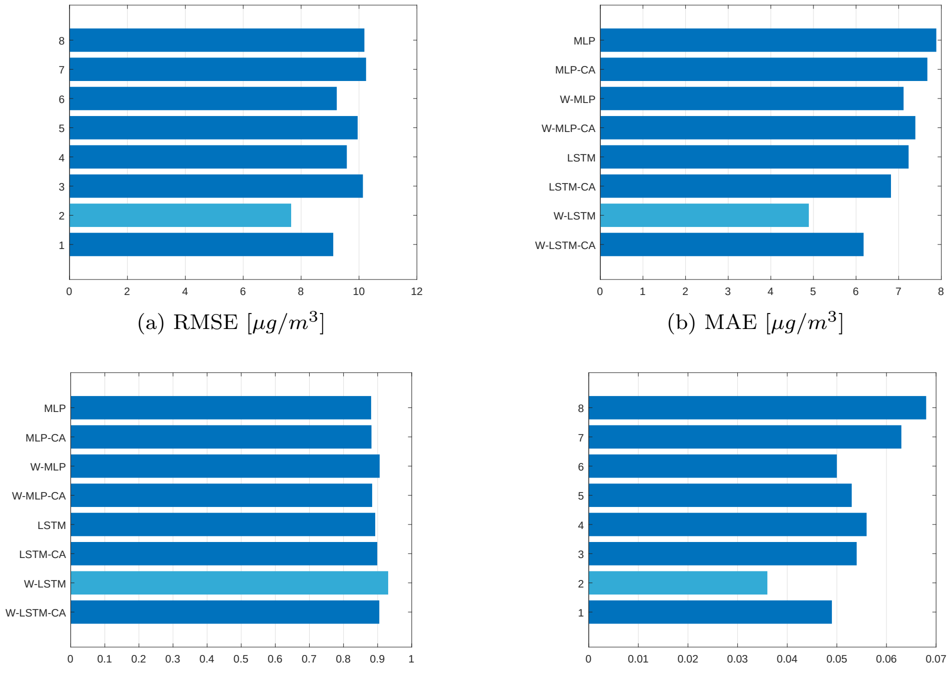
(a) RMSE $[\mu g/m^3]$

(b) MAE $[\mu g/m^3]$

(c) R^2

(d) FB

Figure 15: Forecasting results for CAM site.



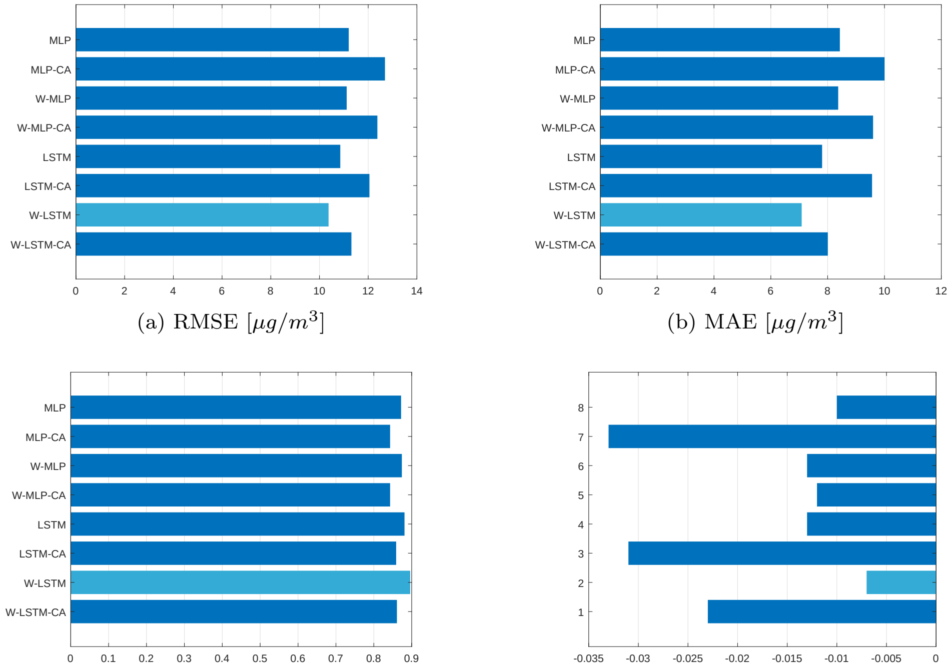
(a) RMSE [$\mu g/m^3$]

(b) MAE [$\mu g/m^3$]

(c) R^2

(d) FB

Figure 16: Forecasting results for BLM site.



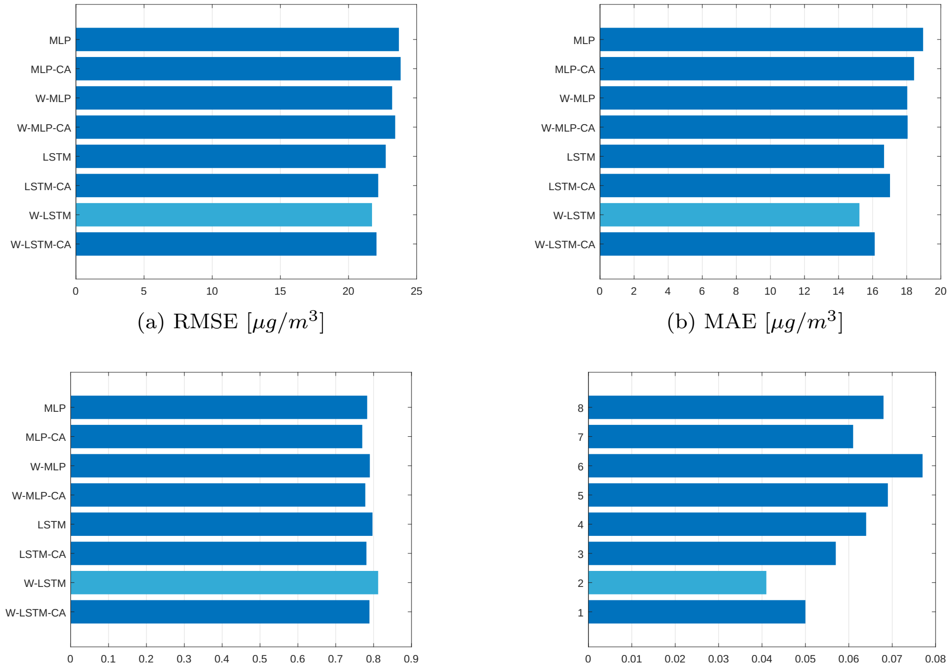
(a) RMSE [$\mu g/m^3$]

(b) MAE [$\mu g/m^3$]

(c) R^2

(d) FB

Figure 17: Forecasting results for WST site.



(a) RMSE [$\mu g/m^3$] (b) MAE [$\mu g/m^3$]

(c) R^2 (d) FB

Figure 18: Forecasting results for HAM site.

5.8. Error distribution and scatter plots of model results

The distribution of the forecasting errors of both LSTM-based and MLP based models are shown in Figure 19 and 20, respectively. Consistent with the results in Table 8, the error distribution of the forecasts of the best-performing model, e.g. the W-LSTM model for BLM site, is centred at 0 $\mu g/m^3$. On the other hand, the error histograms of the results of the worst-performing model, e.g. the W-MLP and MLP models for MAR site, are negatively-skewed. The same histograms also exhibit a wide range of error values, e.g. from around -100 $\mu g/m^3$ to around 50 $\mu g/m^3$.

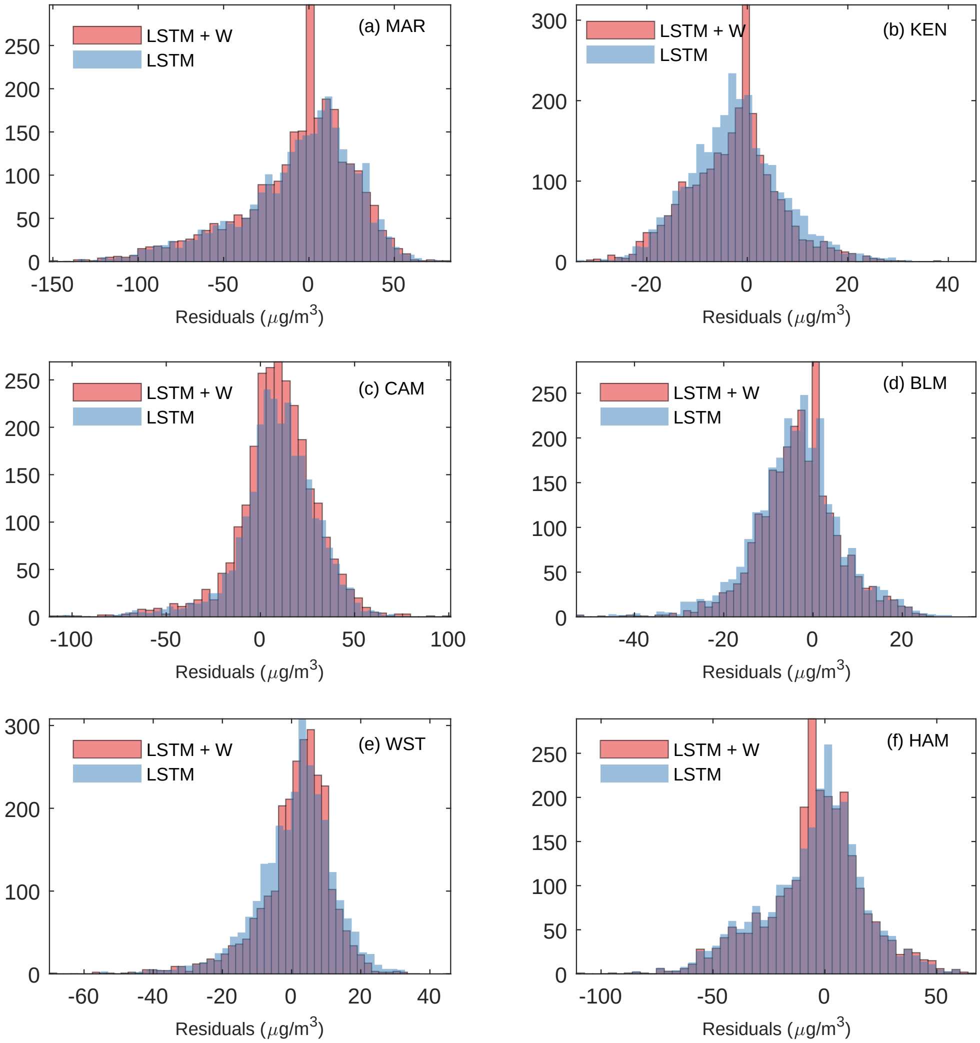


Figure 19: Error histograms of the forecasting results of both proposed and benchmark LSTM models for (a) MAR, (b) KEN, (c) CAM, (d) BLM, (e) WST, and (f) HAM sites.

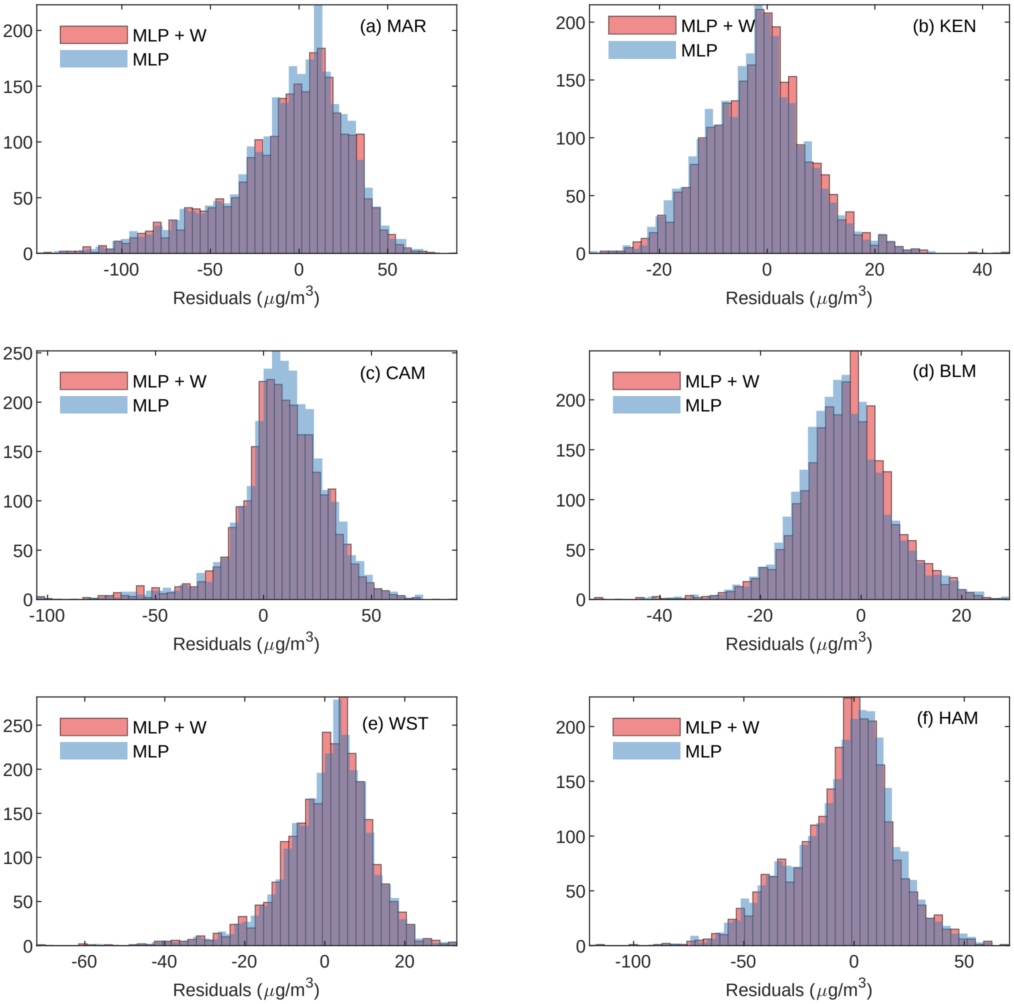


Figure 20: Error histograms of the forecasting results of both proposed and benchmark MLP models for (a) MAR, (b) KEN, (c) CAM, (d) BLM, (e) WST, and (f) HAM sites.

The scatter plots of the results produced by the W-LSTM and W-MLP models are shown in Figures 22 and 21. As shown in Figures 21b, 21d, and 21e, the actual and predicted NO_2 data points from the W-LSTM models for KEN, BLM and WST sites are concentrated near the ideal fit. In addition, the said models demonstrate their ability to capture the higher values of concentrations more accurately. A similar observation can be made for the results of the W-MLP models for KEN, BLM and WST sites (see Figures 22b, 22d, and 22e). In contrast, the poor performance of the models for MAR site is clearly depicted in Figures 22a and 21a. The scatter plot between the actual data and the results of the W-MLP model for MAR site also exhibits a very high tendency to over-and under-predict, e.g. FB = 0.0516. This is also true for the W-MLP model for CAM site which yielded the worst FB score, FB = 0.1173. In general, the plots demonstrate the suitability of the proposed wavelet-based approach for forecasting NO_2 levels at sites utilising only the data from their neighbouring sites.

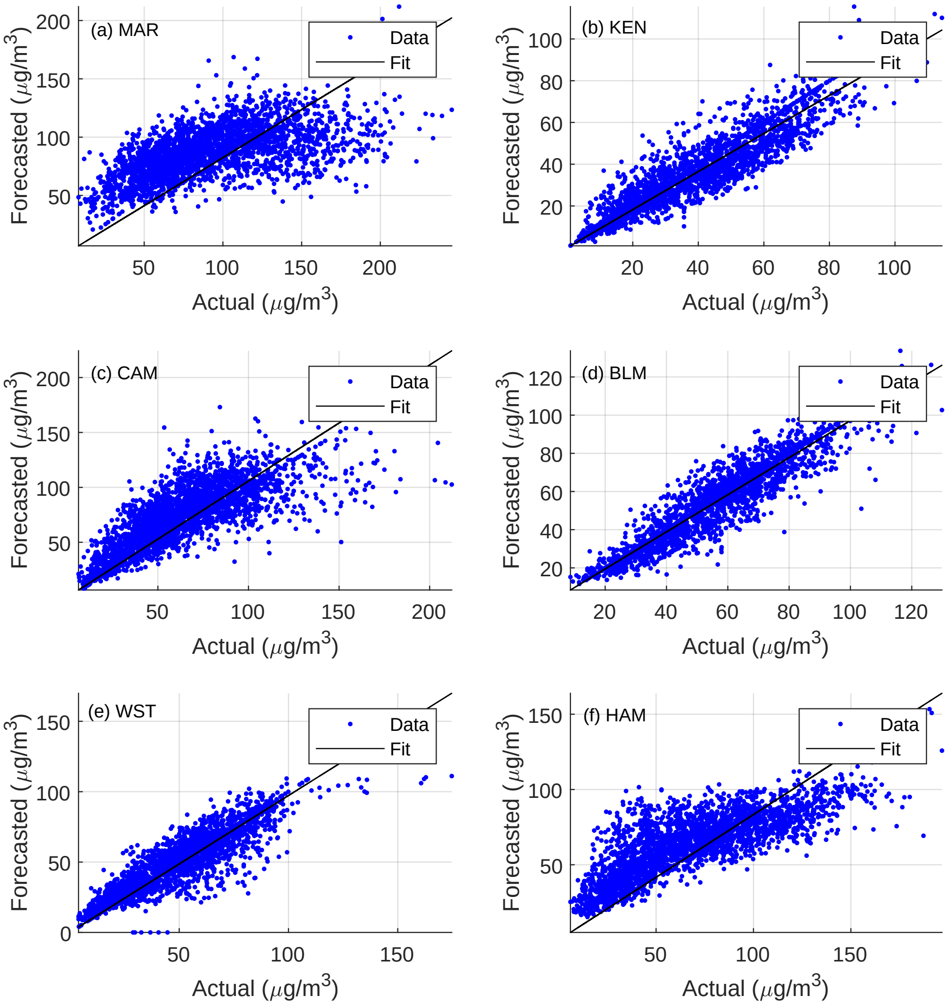


Figure 21: Scatter plots of the actual NO_2 data and the forecasting results of the wavelet-based MLP models for (a) MAR, (b) KEN, (c) CAM, (d) BLM, (e) WST and (f) HAM sites.

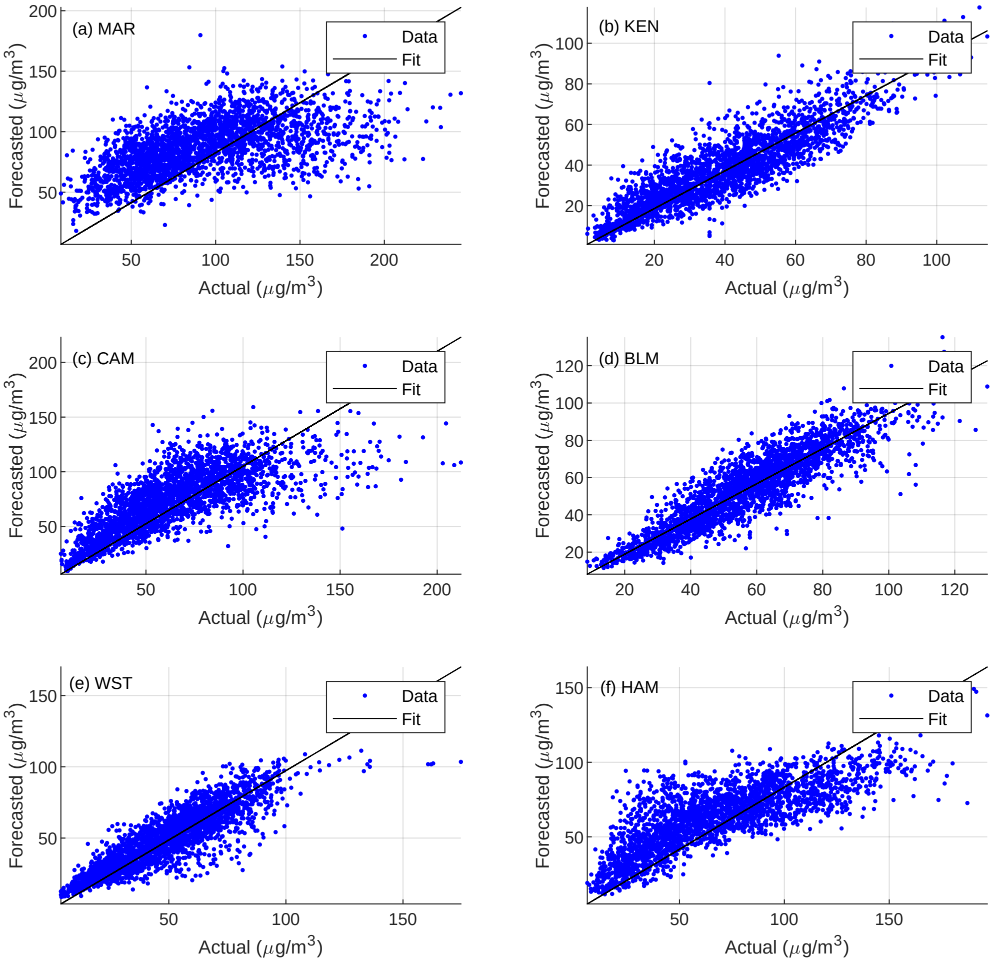


Figure 22: Scatter plots of the actual NO_2 data and the forecasting results of the wavelet-based MLP models for (a) MAR, (b) KEN, (c) CAM, (d) BLM, (e) WST and (f) HAM sites.

5.9. Probability of detection and false alarm rate results

The additional statistical performance metrics results are summarised in Table 9. The W-LSTM and LSTM results for KEN, CAM and WST sites reveal the ability of the models to correctly forecast NO_2 episodes. In general, the wavelet-based LSTM models provided the best results although the wavelet-based MLP models also obtained satisfactory predictions. The results also reveal that the application of DWT improves the ability of the plain LSTM and MLP models to correctly identify peak NO_2 pollution values. Both W-LSTM and W-MLP models for MAR site provided fairly reasonable results despite their generally poor results shown in the previous sections. The frequency of peak hourly NO_2 data from MAR site may have helped the models learn to approximate and reproduce peak values. However, the LSTM and W-LSTM models for both BLM and HAM sites performed poorly. The said models correctly identified only 19% to 22% of the potential high levels of the actual NO_2 concentration data. On the other hand, all models that demonstrated good performance in correctly detecting potential episodes seem to display higher tendencies to issue false alarms. The W-LSTM models for KEN, CAM and WST performed poorly, releasing false alerts 44% to 71% of the time. In contrast, the models for HAM and MAR sites exhibited the least tendencies, e.g. FAR of 0.03% and 19.10%, respectively, in issuing false alarms. In general, the results shown in the table indicate that the wavelet-based deep learning approach improves the overall ability of the models for all sites to correctly identify actual episodes and avoid issuing false alerts.

Table 9: The results of different models for the 1h-ahead forecasting of hourly NO_2 levels.

		Target sites					
Model Code	Alarm Results	MAR	KEN	CAM	BLM	WST	HAM
W-LSTM	No. of predicted episodes	810	9	461	18	20	162
	No. of correctly predicted episodes	716	7	211	9	15	126
	POD	0.612	0.909	0.718	0.188	0.882	0.216
	No. of false alarms	147	4	233	6	10	4
	FAR	0.191	0.444	0.524	0.333	0.714	0.027
LSTM	No. of predicted episodes	740	11	463	15	20	135
	No. of correctly predicted episodes	603	4	198	9	13	82
	POD	0.532	0.801	0.623	0.196	0.565	0.148
	No. of false alarms	159	8	296	7	11	5
	FAR	0.197	0.464	0.614	0.383	0.015	
W-MLP	No. of predicted episodes	721	9	490	16	26	81
	No. of correctly predicted episodes	563	5	186	9	14	73
	POD	0.499	0.649	0.576	0.196	0.609	0.145
	No. of false alarms	168	4	294	7	7	2
	FAR	0.247	0.444	0.6204	0.438	0.350	0.030
MLP	No. of predicted episodes	680	14	471	15	20	67
	No. of correctly predicted episodes	512	5	177	7	13	63
	POD	0.477	0.625	0.543	0.152	0.565	0.129
	No. of false alarms	185	9	304	8	12	8
	FAR	0.257	0.6429	0.6242	0.533	0.462	0.100
	No. of observed episodes	1074	8	326	46	23	503

6. Conclusions

In this paper, a novel hybrid forecasting approach based on deep learning neural networks and discrete wavelet transformation was applied in the 1-h ahead spatial forecasting of hourly NO_2 levels at six urban locations in Central London.

The novelty of this approach is that only the air pollution data from the neighbouring sites were utilised to estimate the air pollution level at a given target site. This approach offers a high theoretical significance of the techniques proposed as other explanatory variables for training the models were not utilised. Another significance of the proposed modelling approach is the decomposition of the original data into subseries based on wavelets at various levels with lesser variability and the individual forecasting of the said subseries to increase the accuracy of the final result. Finally, the proposed approach employs LSTM models to capture the long-term temporal tendency and provide more accurate estimates of the decomposed subseries. Hybrid and plain models based on a feed-forward model were also built to test the effectiveness of the proposed modelling approach.

The numerical results demonstrate the effectiveness of the wavelet-based LSTM models in improving the results of the wavelet-based MLP models, and the plain LSTM and MLP models, despite the extra steps required to perform wavelet transformation. The proposed models are able to account for 77% to 93% of the variance of the actual NO_2 data at almost all sites. The location of the target sites is identified to influence the performance of the developed models. That is, the data collected at kerbside sites exhibit high variability making them difficult to estimate. The level of mutual correlation between the collected data of all the monitoring sites is also identified to affect model results. Finally, the utilisation of fewer predictors from the neighbouring sites influences the performance of the models. A further examination of the use of techniques that can identify the most representative sites to provide sufficient information to the proposed models is therefore needed.

The proposed hybrid deep learning modelling approach has a great potential to be operationally employed in providing air pollution forecasts in areas that lack monitoring or a good database. Although there is no clean-cut approach to building data-driven models such as LSTM models, the techniques and principles applied in developing the models in this paper can be applied to datasets collected from other locations.

Finally, the findings of the study highlight the ability of hybrid deep learning approaches in delivering better performance. There is a need to focus on more sophisticated hybrid modelling techniques, although a trade-off between model complexity and performance should be carefully considered. In cases where the computational resources are restricted and the amount of training data is limited, the development of an effective yet parsimonious model is more ideal.

7. Acknowledgements

The authors would like to thank the British Council and the Commission on Higher Education (CHED) of the Republic of the Philippines (Grant no. 261810845) for funding this research, and UK-AIR, DEFRA (https://uk-air.defra.gov.uk/) for making the air quality data available online.

References

WHO, WHO — Ambient (outdoor) air quality and health, 2016. URL: `http://www.who.int/mediacentre/factsheets/fs313/en/`.

OECD, Policy Highlights - The economic consequences of outdoor air pollution, Technical Report, Organisation for Econonomic Co-operation and Development, 2016.

European Environmental Agency, Air quality in Europe — 2018 report, 5, 2018. URL: `papers2://publication/uuid/1D25F41B-C673-4FDA-AB71-CC5A2AD97FDD`. doi:10.2800/62459.

A. Baklanov, O. Hänninen, L. H. Slørdal, J. Kukkonen, N. Bjergene, B. Fay, S. Finardi, S. C. Hoe, M. Jantunen, A. Karppinen, A. Rasmussen, A. Skouloudis, R. S. Sokhi, J. H.

Sørensen, V. Ødegaard, Integrated systems for forecasting urban meteorology, air pollution and population exposure, Atmos. Chem. Phys. Atmos. Chem. Phys. 7 (2007) 855–874. URL: www.atmos-chem-phys.net/7/855/2007/.

F. A. Gers, D. Eck, J. Schmidhuber, Applying LSTM to Time Series Predictable through Time-Window Approaches, in: G. Dorffner, H. Bischof, K. Hornik (Eds.), Artif. Neural Networks — ICANN 2001, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 669–676.

J. McLaren, I. Williams, The impact of communicating information about air pollution events on public health, Sci. Total Environ. 538 (2015) 478–491. URL: https://linkinghub.elsevier.com/retrieve/pii/S0048969715304903. doi:10.1016/j.scitotenv.2015.07.149.

H. Chen, Q. Li, J. S. Kaufman, J. Wang, R. Copes, Y. Su, T. Benmarhnia, Effect of air quality alerts on human health: a regression discontinuity analysis in Toronto, Canada, Lancet Planet. Heal. 2 (2018) e19–e26. URL: https://www.sciencedirect.com/science/article/pii/S2542519617301857?via%3Dihub. doi:10.1016/S2542-5196(17)30185-7.

S. M. S. Cabaneros, J. K. Calautit, B. R. Hughes, A review of artificial neural network models for ambient air pollution prediction, Environ. Model. Softw. 119 (2019) 285–304. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364815218306352. doi:10.1016/j.envsoft.2019.06.014.

O. Conti, B. Heibati, I. Kloog, M. Fiore, M. Ferrante, A review of airq models and their applications for forecasting the air pollution health outcomes, Environmental science and pollution research international 24 (2017) 6426–6445. URL: https://search.proquest.com/docview/1855790913?accountid=14116.

M. E. Chang, C. Cardelino, Application of the Urban Airshed Model to forecasting next-day peak ozone concentrations in Atlanta, Georgia., J. Air Waste Manag. Assoc. 50 (2000) 2010–2024. URL: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=11111345&site=ehost-liv
doi:10.1080/10473289.2000.10464219.

M. T. Chuang, Y. Zhang, D. Kang, Application of WRF/Chem-MADRID for real-time air quality forecasting over the Southeastern United States, Atmos. Environ. 45 (2011) 6241–6250. URL: http://dx.doi.org/10.1016/j.atmosenv.2011.06.071. doi:10.1016/j.atmosenv.2011.06.071.

Z. Jacobson, Development and Application of a New Air Pollution Modeling System - II. Aerosol Module Structure and Design, Atmos. Environ. 31 (1997) 131–144. doi:10.16011/j.cnki.jjwt.2015.07.004.

A.-L. Dutot, J. Rynkiewicz, F. E. Steiner, J. Rude, A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions, Environ. Model. Softw. 22 (2007) 1261–1269. URL:

https://www.sciencedirect.com/science/article/pii/S1364815206001976. doi:10.1016/J.ENVSOFT.2006.08.002.

National Research Council, Models in environmental regulatory decision making, National Academies Press, 2007.

K. Y. Ng, N. Awang, Multiple linear regression and regression with time series error models in forecasting PM 10 concentrations in Peninsular, Environ. Monit. Assess. (2018). doi:10.1007/s10661-017-6419-z.

S. Q. Dotse, M. I. Petra, L. Dagar, L. C. De Silva, Application of computational intelligence techniques to forecast daily PM10exceedances in Brunei Darussalam, Atmos. Pollut. Res. 9 (2018) 358–368. URL: https://doi.org/10.1016/j.apr.2017.11.004. doi:10.1016/j.apr.2017.11.004.

B.-C. Liu, A. Binaykia, P.-C. Chang, M. Tiwari, C.-C. Tsao, Urban air quality forecasting based on multi- dimensional collaborative Support Vector Regression (SVR): A case study of Beijing- Tianjin-Shijiazhuang, PLoS One 12 (2017) 1–17. URL: http://dx.plos.org/10.1371/journal.pone.0179763. doi:10.1371/journal.pone.0179763.

F. Franceschi, M. Cobo, M. Figueredo, Discovering relationships and forecasting PM10and PM2.5concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering, Atmos. Pollut. Res. (2018) 0–1. URL: http://dx.doi.org/10.1016/j.apr.2018.02.006. doi:10.1016/j.apr.2018.02.006.

M. Gardner, S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmos. Environ. 32 (1998) 2627–2636. URL: http://www.sciencedirect.com/science/article/pii/S1352231097004470. doi:10.1016/S1352-2310(97)00447-0.

H. T. Shahraiyni, S. Sodoudi, Statistical modeling approaches for pm10 prediction in urban areas; A review of 21st-century studies, Atmosphere (Basel). 7 (2016) 10–13. doi:10.3390/atmos7020015.

J. Colls, Air Pollution 2nd Edition, Spon Press, 29 West 35th Street, New York, NY 10001, 2001.

M. S. Alam, A. McNabola, Exploring the modeling of spatiotemporal variations in ambient air pollution within the land use regression framework: Estimation of PM10 concentrations on a daily basis, J. Air Waste Manage. Assoc. 65 (2015) 628–640. URL: http://www.tandfonline.com/doi/full/10.1080/10962247.2015.1006377. doi:10.1080/10962247.2015.1006377.

A. Alimissis, K. Philippopoulos, C. G. Tzanis, D. Deligiorgi, Spatial estimation of urban air pollution with the use of artificial neural network models, Atmos. Environ. (2018). URL: https://doi.org/10.1016/j.atmosenv.2018.07.058. doi:10.1016/j.atmosenv.2018.07.058.

C. G. Tzanis, A. Alimissis, K. Philippopoulos, D. Deligiorgi, Applying linear and nonlinear models for the estimation of particulate matter variability *, Environ. Pollut. (2019). URL: https://doi.org/10.1016/j.envpol.2018.11.080. doi:10.1016/j.envpol.2018.11.080.

K. Siwek, S. Osowski, Engineering Applications of Artificial Intelligence Improving the accuracy of prediction of PM 10 pollution by the wavelet transformation and an ensemble of neural predictors, Eng. Appl. Artif. Intell. 25 (2012) 1246–1258. doi:10.1016/j.engappai.2011.10.013.

X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, J. Wang, Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation, Atmos. Environ. 107 (2015) 118–128. URL: https://www.sciencedirect.com/science/article/pii/S1352231015001491. doi:10.1016/J.ATMOSENV.2015.02.030.

Y. Bai, Y. Li, X. Wang, J. Xie, C. Li, Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions, Atmos. Pollut. Res. 7 (2016) 557–566. URL: http://dx.doi.org/10.1016/j.apr.2016.01.004. doi:10.1016/j.apr.2016.01.004.

S. Osowski, K. Garanty, Forecasting of the daily meteorological pollution using wavelets and support vector machine, Eng. Appl. Artif. Intell. 20 (2007) 745–755. doi:10.1016/j.engappai.2006.10.008.

B. S. Freeman, G. Taylor, B. Gharabaghi, J. Thé, Forecasting air quality time series using deep learning, J. Air Waste Manage. Assoc. 68 (2018) 866–886. URL: https://doi.org/10.1080/10962247.2018.1459956. doi:10.1080/10962247.2018.1459956.

X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, T. Chi, Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation, Environ. Pollut. 231 (2017) 997–1004. URL: https://www.sciencedirect.com/science/article/pii/S0269749117307534. doi:10.1016/J.ENVPOL.2017.08.114.

Q. Wu, H. Lin, A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors, Sci. Total Environ. 683 (2019) 808–821. URL: https://doi.org/10.1016/j.scitotenv.2019.05.288. doi:10.1016/j.scitotenv.2019.05.288.

C. Li, Z. Zhu, Research and application of a novel hybrid air quality early-warning system: A case study in China, Sci. Total Environ. 626 (2018) 1421–1438. URL: https://doi.org/10.1016/j.scitotenv.2018.01.195. doi:10.1016/j.scitotenv.2018.01.195.

J. Zhao, F. Deng, Y. Cai, J. Chen, Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction, Chemosphere (2019). URL: https://doi.org/10.1016/j.chemosphere.2018.12.128. doi:10.1016/j.chemosphere.2018.12.128.

J. Ma, Y. Ding, V. J. L. Gan, C. Lin, Z. Wan, Spatiotemporal Prediction of PM2.5 Concentrations at Different Time Granularities Using IDW-BLSTM, IEEE Access 7 (2019) 107897–107907. URL: https://ieeexplore.ieee.org/document/8784234/. doi:10.1109/ACCESS.2019.2932445.

J. Wang, G. Song, A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction, Neurocomputing 314 (2018) 198–206. URL: https://www.sciencedirect.com/science/article/pii/S0925231218307859. doi:10.1016/J.NEUCOM.2018.06.049.

King's College London, London Air Quality Network - King's College London, 2019. URL: http://www.londonair.org.uk/LondonAir/Default.aspx.

DEFRA, AIR QUALITY EXPERT GROUP Nitrogen Dioxide in the United Kingdom, 2004. URL: www.defra.gov.uk/environment/airquality.

WHO, Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide Report on a WHO Working Group OZONE-adverse effects NITROGEN DIOXIDE-adverse effects AIR POLLUTANTS, ENVIRONMENTAL-adverse effects META-ANALYSIS AIR-standards GUIDELINES, Technical Report, World Health Organization, 2003. URL: http://www.euro.who.int/__data/assets/pdf_file/0005/112199/E79097.pdf.

S. G. Mallat, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693. doi:10.1109/34.192463.

Y. Nievergelt, Wavelets Made Easy, reprint of the 2001 edition ed., Birkhauser, NY, 2013.

I. Daubechies, Ten lectures on wavelets., SIAM Press, Philadelphia, USA., 1988.

D. Dunea, A. Pohoata, S. Iordache, Using wavelet–feedforward neural networks to improve air pollution forecasting in urban environments, Environ. Monit. Assess. 187 (2015). doi:10.1007/s10661-015-4697-x.

W. S. McCulloch, W. Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics 5 (1943) 115–116. doi:https://doi.org/10.1007/BF02478259.

C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Inc., New York, NY, USA, 1995.

M. T. Hagan, H. B. Demuth, M. H. Beale, Neural Network Design, Bost. Massachusetts PWS 2 (1995) 734. URL: http://books.google.ru/books?id=bUNJAAAACAAJ%5Cnhttp://ecee.colorado.edu/academics/sch doi:10.1007/1-84628-303-5.

S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735. arXiv:1206.2944.

M. A. Tanner, W. H. Wong, The calculation of posterior distributions by data augmentation, Journal of the American Statistical Association 82 (1987) 528–540. URL: http://www.jstor.org/stable/2289457.

K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359–366. doi:10.1016/0893-6080(89)90020-8. arXiv:arXiv:1011.1669v3.

D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv e-prints (2014) arXiv:1412.6980. arXiv:1412.6980.

K. Levenberg, A METHOD FOR THE SOLUTION OF CERTAIN NON-LINEAR PROBLEMS IN LEAST SQUARES, Technical Report 2, Brown University, 1944. URL: https://www-jstor-org.sheffield.idm.oclc.org/stable/pdf/43633451.pdf?refreqid=excelsio

D. W. Marquardt, An Algorithm for Least-Squares Estimation of Nonlinear Parameters, Technical Report 2, Society for Industrial and Applied Mathematics, 1963. URL: https://www-jstor-org.sheffield.idm.oclc.org/stable/pdf/2098941.pdf?refreqid=excelsior

USEPA, AERMOD Implementation Guide, 2016. URL: https://www3.epa.gov/ttn/scram/models/aermod/aermod_implementation_guide.pdf.

I. The MathWorks, MATLAB Documentation - MathWorks United Kingdom, 2019. URL: https://uk.mathworks.com/help/matlab/index.html.