Neural network-based meta-modelling approach for estimating spatial distribution of air pollutant levels

H. Wahid[a,b], Q.P. Ha[b,*], H. Duc[c], M. Azzi[d]

[a] *Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia*
[b] *Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway, NSW 2007, Australia*
[c] *Office of Environment and Heritage, PO Box 29, Lidcombe, NSW 1825, Australia*
[d] *CSIRO Energy Technology, PO Box 52, North Ryde, NSW 1670, Australia*

Abstract

Continuous measurements of the air pollutant concentrations at monitoring stations serve as a reliable basis for air quality regulations. Their availability is however limited only at locations of interest. In most situations, the spatial distribution beyond these locations still remains uncertain as it is highly influenced by other factors such as emission sources, meteorological effects, dispersion and topographical conditions. To overcome this issue, a larger number of monitoring stations could be installed, but it would involve a high investment cost. An alternative solution is via the use of a deterministic air quality model (DAQM), which is mostly adopted by regulatory authorities for prediction in the temporal and spatial domain as well as for policy scenario development. Nevertheless, the results obtained from a model are subject to some uncertainties and it requires, in general, a significant computation time. In this work, a meta-modelling approach based on neural network evaluation is proposed to improve the estimated spatial distribution of the pollutant concentrations. From a dispersion model, it is suggested that the spatially-distributed pollutant levels (i.e. ozone, in this study) across a region under consideration is a function of the grid coordinates, topographical information, solar radiation and the pollutants precursor emission. Initially, for training the model, the input-output relationship is extracted from a photochemical dispersion model called The Air Pollution Model and Chemical Transport Model (TAPM-CTM), and some of those input-output data are correlated with the ambient measurements collected at monitoring stations. Here, improved radial basis function networks, incorporating a proposed technique for selection of the network centres, will be developed and trained by using the data obtained and the forward selection approach. The methodology is then applied to estimate the ozone concentrations in the Sydney basin, Australia. Once executed, apart from the advantage of inexpensive computation, it provides more reliable results of the estimation and offers better predictions of ozone concentrations than those obtained by using the TAPM-CTM model only, when compared to the measurement data collected at monitoring stations.

Keywords: Metamodel; spatial distribution; ozone; Radial basis function networks; TAPM-CTM

1. Introduction

As cities and their surrounding suburbs around the world expand with increasing people, motor vehicles and industries, there is an urgent need to understand the connection between air pollution formation, human health, and emission control with urban management. Since quality air is associated with healthy society and clean environment, the accurate assessment of the air pollutant levels is an important task for the authorities to determine appropriate management environmental policies. In general, air quality assessment can be conducted using three different staged approaches; air quality monitoring, emission inventory and assessment, and air quality modelling. Each has its own usefulness to the policy maker for understanding the air pollution nature due to various sources in the urban setting, in both temporal and spatial aspects.

The spatial distribution estimation of air pollutants using data measurement is usually limited by the number of available monitoring stations across a region. To tackle this problem, one way [1] is via the use of mobile measurement stations, that are movable to other locations after some period of time to avoid expensive investments by increasing the number of fixed monitoring sites. However, this is generally difficult to be implemented, time-consuming and unlikely to be accessible at most of rural locations. Air quality models could also be used for a more cost effective method [2, 3, 4]. Nevertheless, their simulation results are much dependent on the correct formulation of chemical reactions involved in the models as well as the accuracy of emission inventory data and meteorological data used as inputs. Furthermore, air quality models also imply a high computational cost, which generally require several days or weeks for a particular simulation task, depending on the model and the problem in consideration.

Thus, to reduce the computation burden for simulation, appropriate and reliable statistical techniques could be implemented. For example, Duc et al. [5] used a Kriging approach to study the spatial correlation of SO_2, NO, NO_2 and ozone (O_3) over a long-distance network in Sydney, Australia. They found that

*Corresponding author. Tel.: +61 2 9514 2453; fax: +61 2 9514 2868
Email address: Quang.Ha@uts.edu.au (Q.P. Ha)

within a 30km radius, this method showed a reasonable correlation for some air pollutants, but not likely for ozone due to the non-linearity and complicacy of its formation. Soft computing based on artificial intelligence (AI) can serve as an alternative in environmental science studies. In climate control, Trabelsi et al. [6] implemented a fuzzy clustering technique to model air temperature and humidity inside a greenhouse to increase the crop production. More recently, Fazel Zarandi et al. [7] used the type-2 fuzzy logic theory to construct a model for the prediction of carbon monoxide in Tehran, Iran. A comparative analysis on statistical approaches for ozone prediction has been conducted in [8]. It is found that among neural networks (NN), support vector regression (SVR) methods and those with uncertainty, models of SVR with polynomial kernel functions appeared to perform better than neural networks (feed-forward NN, time delay NN, and RBFNN) in terms of the root mean squared error (*RMSE*). However, their attempt in ozone predictions is actually similar to previous authors (e.g. [9, 10, 11]), where input parameters for training are chosen from available measured air pollutant and meteorological data without taking into account the spatial distribution of the pollutants.

In air quality research, neural networks have been successfully applied to model some air quality predictions, mainly in forecasting the pollutant concentration (i.e. temporal predictions), see e.g. [12, 13, 14, 15]. An air dispersion model and neural networks were integrated to reduce the complexity of the spatial predictions in the simulation of complex situations [16], but without improving reliability via verification with measurement data collected. Carnevale et al. [17] proposed neural network models to estimate a non-linear source-receptor relationship for ozone and PM_{10} concentrations, where the networks were trained from input-output data generated by a deterministic model. Good results for the pollutions mapping were shown therein as compared to the deterministic model, again without validating with results obtained from the actual sites' measurements. Moreover, meteorological data were not considered as the model input. Pfeiffer et al. [18] used diffusive sampling measurements and neural networks to compute the average spatial distribution of air pollutants in Cyprus. However, a large number of the diffusive samplers is required to get the correct spatial map for a particular pollutant: 270 samplers are needed at 270 sites for NO_2 pollutant.

To enhance the prediction performance for the spatial estimation of air pollutant profiles, we propose the integration of three approaches in the modelling, i.e. deterministic air quality model, neural network model and ambient measurement data. With this, we aim to estimate, with high accuracy, the spatial distribution of the ozone, as an air pollutant, across a region. A number of estimated pollutant levels of interest has been computed such as the 1-hour, 4-hour, 8-hour, or 24-hour daily maximum average by using a radial basis function neural network (RBFNN) metamodel with an improved algorithm to select the network centres. Here, a deterministic model, The Air Pollution Model and Chemical Transport Model (TAPM-CTM), developed by the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO), is used to produce a modelled output in which some important grid data values in the model region are extracted to become inputs and outputs of the neural network-based metamodel. These values have been post-processed to correlate with the ambient measurement data. On the other hand, to enhance the reliability of the precursor emission data of ozone (from an inventory database), a Gaussian dispersion model is used to transform the measured precursor's concentration data at monitoring stations to become additional emission data. The effectiveness of the model is then determined through some performance indices, and the results will be verified with measurement data from other sites, which have not been used in the training process. As the conceptual framework of the approach is generic, the proposed implementation can be extended for the estimation of other air pollutants for their temporal and spatial distributions.

The paper is organised as follows. After the introduction, Section 2 describes the proposed RBFNN metamodel together with a new technique for selection of the network centres. The estimation of the spatial distribution of the air pollutant is discussed in Section 3. Section 4 presents the results across a region and discussion for this case study. Finally, Section 5 gives some concluding remarks.

2. Radial basis function network metamodel

For describing characteristics and behaviour of very complex systems, the discrete event model approach has offered good estimation accuracy but may suffer from a difficulty on the realisation and the high demand of computational expenses. Therefore, metamodels have been suggested to be an approximate model that can adequately represent the intrinsically non-linear and complex relationship between the systems input and output. Splines, neural networks, kriging and support vector machine are some of the proposed methods in the literature for metamodels [19, 20]. To this end, the radial basis function neural network (RBFNN) can offer good performance on accuracy, robustness, problem types, sample size, efficiency, and simplicity as compared to stochastic approaches [21, 22]. Due to these advantages, RBFNN has attracted many researchers in various real-life applications, see e.g., [23, 24].

In the RBFNN, three difficulties involved in the training algorithm include the selection of the radial basis centres, of the basis function radius (spread), and of network weights. For the choice of network centres, several methods have appeared in the literature, which can be grouped to random, unsupervised and supervised selection (see [25, 26, 27]). Also known as the forward selection [28], the supervised selection is a systematic way utilising the orthogonal least square algorithm. In this work, we use the hidden neuron output information from a previous iteration. The idea is partially adopted from the forward selection method by Orr [29] in conjunction with the weighted least squares (WLS) theory, which gives the advantage in dealing with noisy data.

The RBFNN output vector, of dimension m , corresponding to the input vector $\mathrm{x} \in \Re^n$ is mathematically represented as

follows:

$$f_i(x) = \sum_{k=1}^{n} w_{ki} b(\|\mathrm{x} - c_k\|_2), \quad i = 1,2,...,m \tag{1}$$

where $b(\cdot)$ is a basis function, $\|\cdot\|_2$ denotes the Euclidean norm, w_{ki} are weights in the output layer, $c_k \in \Re^n$ are the RBF centres in the input vector space, and n is the number of neurons (and centres) in the hidden layer. In matrix notation, equation (1) can also be written as:

$$F = B^T W, \tag{2}$$

where F is the matrix of the network output with $p \times m$ dimension, B is the matrix of hidden nodes with $n \times p$ dimension, $W = [w_{ki}]^T$ is a network weight matrix with $n \times m$ dimension, and p is the number of dataset patterns.

By incorporating the WLS theory, the RBF weights can be computed by the following equation,

$$W_{RBF} = (BHB^T)^{-1} BHD, \tag{3}$$

where D is the $p \times m$ matrix of the desired output and H is the diagonal matrix of the least square weighting coefficients with diagonal components h_{jj}, where $1 < j < p$.

From (3), the weights at the k-th iteration can be trained by the following equation,

$$W_k = (B_k H_k {B_k}^T)^{-1} B_k H_k D = (A_k)^{-1} B_k H_k D, \tag{4}$$

where the variance matrix $A_k = B_k H_k {B_k}^T$ can be formed as:

$$\begin{aligned} A_k &= \begin{bmatrix} B_{k-1} \\ {b_k}^T \end{bmatrix} [H_k] \begin{bmatrix} {B_{k-1}}^T & b_k \end{bmatrix} \\ &= \begin{bmatrix} B_{k-1} H_k {B_{k-1}}^T & B_{k-1} H_k b_k \\ {b_k}^T H_k {B_{k-1}}^T & {b_k}^T H_k b_k \end{bmatrix} \\ &= \begin{bmatrix} A_{k-1} & B_{k-1} H_k b_k \\ {b_k}^T H_k {B_{k-1}}^T & {b_k}^T H_k b_k \end{bmatrix}, \end{aligned} \tag{5}$$

and the inverse matrix of (5) is given by:

$$\begin{aligned} {A_k}^{-1} &= \frac{1}{det(A_k)} \begin{bmatrix} {b_k}^T H_k b_k & -B_{k-1} H_k b_k \\ -{b_k}^T H_k {B_{k-1}}^T & A_{k-1} \end{bmatrix} \\ &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \end{aligned} \tag{6}$$

Here, we are concerned with a direct relationship between A_k and A_{k-1} matrices, thus only the first matrix entry in (6) is taken into account. From $A_{k-1} = B_{k-1} H_{k-1} {B_{k-1}}^T$, we can write matrix A_{11} as $A_{11} = (A_{k-1} - ({b_k}^T H_k b_k)^{-1} B_{k-1} H_k b_k {b_k}^T H_k {B_{k-1}}^T)^{-1}$. Therefore, we have

$$A_{k(A_{11})} = A_{k-1} - B_{k-1} H_k b_k ({b_k}^T H_k b_k)^{-1} {b_k}^T H_k {B_{k-1}}^T. \tag{7}$$

By using the small rank adjustment [30], we can obtain

$$\begin{aligned} {A_k}^{-1} = {A_{k-1}}^{-1} + {A_{k-1}}^{-1} q_k ({q_k}^T {A_{k-1}}^{-1} q_k + \\ {b_k}^T H_k b_k)^{-1} {q_k}^T {A_{k-1}}^{-1}, \end{aligned} \tag{8}$$

where $q_k = B_{k-1} H_k b_k$ and $q_k^T = b_k^T H_k {B_{k-1}}^T$. The RBF network output over the training set is given by [31]:

$$F_k = B_k^T W_k = B_k^T A_k^{-1} B_k H_k D. \tag{9}$$

Now, we can estimate the sum of squared error ε_k at the k-th iteration as follows:

$$\varepsilon_k = tr\{(D - F_k)^T H_k (D - F_k)\}, \tag{10}$$

or in a more compact form,

$$\varepsilon_k = tr\{D^T H_k Q_k D\}, \tag{11}$$

where

$$Q_k = I_R - B_k^T A_k^{-1} B_k H_k \tag{12}$$

is a projection matrix, I_R is the identity matrix with the dimension of $p \times p$ and $tr(.)$ is the trace function which computes the sum of the elements in the main diagonal. Using equation (8), matrix Q_k can be re-written as follows:

$$\begin{aligned} Q_k = I_R - B_k^T A_{k-1}^{-1} B_k H_k - B_k^T A_{k-1}^{-1} q_k (q_k^T A_{k-1}^{-1} q_k + b_k^T H_k b_k)^{-1} \\ q_k^T A_{k-1}^{-1} B_k H_k. \end{aligned} \tag{13}$$

Substituting matrix B_k, B_k^T and $H_k = H_{k-1}$, where A_{11} is used for A_k^{-1}, into (13) yields:

$$Q_k = Q_{k-1} - \frac{Q_{k-1} H_k b_k b_k^T H_k Q_{k-1} H_k}{b_k^T H_k Q_{k-1} H_k b_k + b_k^T H_k b_k}, \tag{14}$$

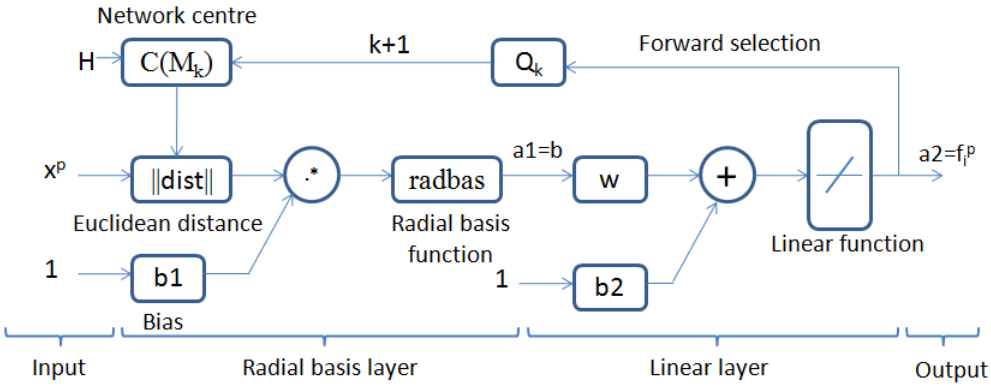

Figure 1: Radial basis function network scheme with forward selection and weighted least square (FSWLS).

in which the denominator part (i.e. $b_k^T H_k Q_{k-1} H_k b_k + b_k^T H_k b_k$) always returns a scalar number. Thus, by implementing equation (11), the error can be calculated as follows:

$$\varepsilon_k = \varepsilon_{k-1} - \frac{tr\{D^T H_k Q_{k-1} H_k b_k b_k^T H_k Q_{k-1} H_k D\}}{b_k^T H_k Q_{k-1} H_k b_k + b_k^T H_k b_k}. \quad (15)$$

This also means that we can minimise the error by maximising M_k given by:

$$M_k = \frac{\|D^T H_k Q_{k-1} H_k b_k\|^2}{b_k^T H_k Q_{k-1} H_k b_k + b_k^T H_k b_k}. \quad (16)$$

To simplify the solution, a constant h_{jj} is considered for elements of the diagonal matrix H_k at any k-th iteration. In other words, H_k can be written as:

$$H_k = \alpha * I_R. \quad (17)$$

The weight coefficient value α, initially set at 1 in [31], is suggested here to lie within the band $\pm$ 25% of its unity value, so as maintain the best convergence region and to avoid the over-fitting problem.

Accordingly, the selection of the network centre can proceed by taking the vector number from a finite set (i.e. iterated evaluation of different vectors b_k) of possible centres corresponding to the maximum value of M_k. However, this procedure may again cause an ill-conditioned problem which hinders the advantage of RBFNN training. To avoid the iteration process and the over-fitting problem, one solution is to choose a smaller number of centres than the dimension of the input space [25]. Thus, we suggest that the set of possible centres can be assessed by the Gram matrix P, as suggested in [32], where P remains a symmetrical and orthogonal matrix of all the possible radial basis output of a given training data. Thus, equation (16) can be re-written as,

$$M_k = \frac{\|D^T H_k Q_{k-1} H_k P\|^2}{sum\,(P H_k Q_{k-1} H_k P + P^T H_k P)}, \quad (18)$$

where $sum(\cdot)$ returns the sum of the values of each matrix column. To save the memory for computation, (18) can be further simplified for faster computation as follows:

$$M_k = \frac{\|D^T H_k Q_{k-1} H_k P\|^2}{sum(P H_k) Q_{k-1} (H_k P) + sum(P^T) H_k P}. \quad (19)$$

To execute the algorithm, at $k = 1$, matrix Q_0 is set as I_R , and at the following iterations Q_{k-1} is set as Q_k, which has been computed in the previous node (i.e. $k-1$) by using (13). A computational algorithm for the proposed RBFNN has been preliminarily reported in [31]. The overall improved network scheme is depicted in Figure 1, wherein the network centre C at the k-th iteration is a function of M_k.

3. Spatial distribution model for air pollutant estimation

3.1. Overview of air quality prediction

Our objective is to construct a model for the spatial prediction of ground level ozone concentrations over a certain large region, i.e. in this case, the Sydney basin in Australia [33]. Notably, the surface ozone is one of the most important photochemical pollutants that require to be controlled because of its impacts on human health and on the environment, as reported in [34].

Compared to the other air pollutants (e.g., sulfur dioxide, carbon monoxide, particulate matters etc.), the ozone formulation is quite complex and non-linear, making it difficult to be predicted. It is typically formed by nitrogen oxides (NO_x) and volatile organic compounds (VOCs) in the presence of solar radiation, and it may cause several negative impacts to the human, vegetation as well as to the environment, at the ground level. Thus, reliable prediction of its level may provide an indication to implement the long-term plan for improving health conditions to the community.

Intensive research and development on the air pollutants prediction tools have been started since the last two decades. The methods can be categorized into two types of approaches; deterministic and statistical models. A spatial distribution estimation usually uses the former type, also known as a dispersion model. It simulates the atmosphere for a certain region by dividing it into a large number of individual grid cells, and estimate pollutant concentrations in each cell by considering the air dispersion effects of pollutants into each cell, the upward and downward movement of the pollutants across an assumed number of atmospheric layers and the amount of emission from many different sources. However, because of its complexity, their execution is quite time consuming, depending on the model used and the scale of the region under consideration. Popular models reported in the literature are CAMx [35], CMAQ [36] and GEOS-CHEM [37].

For the statistical models, most approaches such as regression analysis, interpolation and artificial intelligence, use ambient measurement data. For the spatial distribution estimation, the interpolation algorithms have been used, e.g., kriging in [5]; a local weight function in [38]. However, this methodology only gives rough visualisation to interpolate the measurement results from the monitoring sites, without considering other possible factors such as geographical topology and meteorological conditions.

The artificial intelligence approach is basically effective to be used for the local estimation at monitoring sites and nearby areas only. Of interest are recent works by Carnevale et al. [17] and Pfeiffer et al. [18], using artificial neural networks for spatial estimation of pollutants' concentrations.

3.2. Neural network model development for ozone distribution

3.2.1. Input-output parameters

A neural network model is considered as a black box for mapping the best relationship between the inputs and the outputs of the dataset without knowing the underlying physics of the system. In this work, an improved RBFNN is proposed for the modelling where suitable inputs parameters were selected to get the best possible network configuration. To this end, we utilized specific ambient measurement data and also input-output data from the deterministic air quality model, to train the

RBFNN. In this work, we adopted a specialized DAQM model called as TAPM-CTM, a typical model used for air quality regulatory in Australia.

Since ozone is the pollutant to be considered in this paper, the most related input parameters for training the model are the ozones precursors, the $x-y$ coordinates, the topography information and the solar radiation levels. Basically, there are two important classes of precursors involved in the formation of ozone, namely volatile organic compounds (VOCs) and NO_x. However, VOCs are apparently very difficult to measure, hence VOC data are fully based on the emission rate data extracted from the emission inventory system, whereby the NO_x data could be enhanced by incorporating its measurement data collected at the monitoring stations.

The $x-y$ coordinates represent the cells location (in km) in x and y directions, which normally form a group of $2km \times 2km$ domain cells. By using statistical modeling, the coordinate information is adequate for quick interpolation of measurements between the monitoring stations, but it is not quite accurate, especially for a large distance between sites. To improve the estimation, topography information is added, consisting of the height above the sea level (in m) at each domain cell.

Here, ambient temperature data are used to represent, at each cell, the solar radiation level, which basically is a good indicator proxy variable to the formation of ozone and has a strong correlation to the ozone concentration. Generally, a temperature dataset could be made available from a local meteorological institution such as the Bureau of Meteorology for the Sydney region. The lowest layer data (about $20m$ above the sea level) are also considered. These datasets need to be post-processed as daily maximum temperatures, taken from the daylight hourly temperature, as to represent the activeness of the daily ozone production.

The network output consists of daily 8-hour maximum averaged of the ozone concentration (in part per billion, ppb), which is extracted from the DAQM simulation output. The 8-hour average is selected here in this paper as a demonstration of the approach. The 4-hour or 1-hour can be analysed similarly. As for the ozone predictions, the simulation is only run for the summer months (i.e. December, January and February, in Australia), during which the formation of ozone is most intense. To correlate with the actual measurement data, this dataset is calibrated via regression by analysing the correlation ratio between DAQM output and actual concentration data at all available monitoring sites, for each recorded day. This correlation ratio is then multiplied to the entire cell parameters in the simulated domain. For illustration, the topology of the model network is shown in Figure 2. Finally, the entire inputs and output are normalised (e.g. in the interval between 0 and 1, using 'mapminmax' function in Matlab), in order for them to contribute with the same influence to the RBFNN.

3.2.2. NO_x emission distribution

Generally, the amount of the daily NO_x emission (in kg/day) taken from the emission inventory does not change much for each day, except there is a small difference between the weekdays and the weekend days. Thus, the daily emission can be assumed to be identical over time at one location, however, they are apparently different between each domain cell. To make the significant variations of daily emission for the purpose of neural network training, the actual measured NO_x concentration at monitoring stations (typically in $pphm$) is converted to an emission rate, distributed to the entire domain and added to the original emission data. This can be done by assuming the emission source is at ground level and thus, the produced concentration is contaminated at the ground level and using the basic Gaussian dispersion model developed by Pasquill [39], i.e.

$$C(X,Y,Z) = \frac{Q}{2\pi u \sigma_y(X)\sigma_z(X)} * \exp\left[-\frac{1}{2}\left[\left(\frac{Y}{\sigma_y(X)}\right)^2 + \left(\frac{Z}{\sigma_z(X)}\right)^2\right]\right], \quad (20)$$

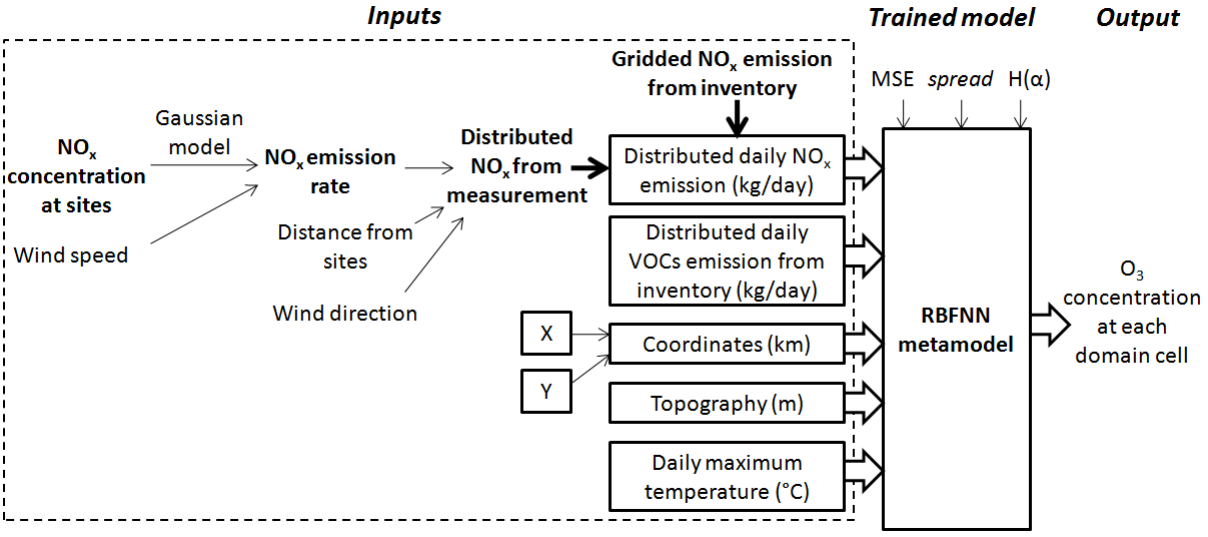

Figure 2: Inputs and output for training the RBFNN metamodel.

where $C(X,Y,Z)$ is the pollutants concentration (in $\mu g/m^3$) at distance X downwind from the source (in *meters*), distance Y crosswind (equal to 0 for this case), and vertical direction Z; Q is the emission rate (in g/sec), u is the average wind speed (in m/sec), σ_y and σ_z are the dispersion coefficient respectively in Y- and Z-direction. The values of σ_y and σ_z have been determined empirically by plume studies available in the literature. They depend on many variables, and especially on the stability of the atmosphere, which is normally rated from A to F, with A being the least stable and F the most stable of an environment. For instance, sigma values can be determined roughly from the dispersion coefficient graphs, or more accurately determined by the following equations [40]:

$$\sigma_y = aX^{0.894}, \tag{21}$$

$$\sigma_z = cX^d - f, \tag{22}$$

where values of a, c, d and f can be obtained by curve-fitting, depending on the atmospheres stability condition. Note that the measurement unit, in *pphm*, for pollutants concentration is consistently converted to $\mu g/m^3$ using the molecular mass of NO and NO_2 at $25^{\circ}C$ and 1 *atm*.

The emission rates, assumed to be coarsely distributed to other cells, are estimated at these cells by considering the nearest distance to the station, adjusted by the wind direction factor. Finally, the calculated distributed NO_x emission is added to the gridded emission rate from the inventory database.

3.2.3. Training, validation and verification

To start the modeling process, firstly we need to define the frame area for the simulation. The border of the domain is approximately selected about 30*km* distance from the most outer monitoring stations for a reasonable correlation.

For the network training purpose, the entire domain is divided to groups of $6km \times 6km$ grid cells for the input dataset from these groups to be able to represent the behavior of the whole frame. This choice reduces the number of datasets to be trained. The dataset was trained by using RBFNN with the appropriate selection of spread parameter (*sp*), least square weighting coefficient (α), and prescribed error goal (*MSE*).

In the validation stage, the denser input-output dataset (i.e. smaller cell size, for e.g. $2km \times 2km$) from the same simulation is used to confirm the correctness of the trained model. The developed model is then tested with other datasets which have not been used in the training stage to predict the spatial distribution of ozone concentration, and the results are compared with the measured ozone level collected at the continuous monitoring sites.

4. Results for case study: Sydney region

4.1. The application domain and the measurement data

The methodology has been applied to the Sydney basin in New South Wales, Australia. The Sydney basin area can be divided into four main regions; East, North West, West and South West based on geographical population settlement pattern. The basin currently has 14 monitoring stations scattered throughout the Sydney metropolitan region, from the coastal area in the East to the edge of the Blue Mountain in the North West and West. Most of the measuring sites are located in the urban area except for some locations, which can be considered as suburban in the greater West, and semi-rural area in the North West.

The whole Sydney region covers the area of about 24,242 km^2. For the station location, in order to get reasonable prediction results using the proposed methodology, the selected domain begins from 246*km* to 384*km* easting and from 6207*km* to 6305*km* northing, by using the Australian map grid (AMG) coordinates, as illustrated in Figure 3.

In this study, the concentrations of two measured air pollutants, ozone (O_3) and nitrogen oxides (NO_x), were measured in part per hundred million (*pphm*) units on an hourly basis. The ozone data were measured using the Ecotech Ozone Monitor 9810, which is based on the ultraviolet spectroscopy principle, while the nitrogen oxides (NO_x) were measured using the Ecotech 9841 instrument. They were calibrated daily and checked frequently.

4.2. Implementation of neural network metamodel

The model development is based on the ambient measurement of pollutant data, meteorological data and primary or precursor pollutant emission sources data for the year of 2004, considered in this paper as the base year of this study. For preparing the output dataset, few simulations for summer days in 2004 were performed by using the TAPM-CTM model. The TAPM model is a three-dimensional prognostic meteorological and air pollution model, which was developed by the Commonwealth

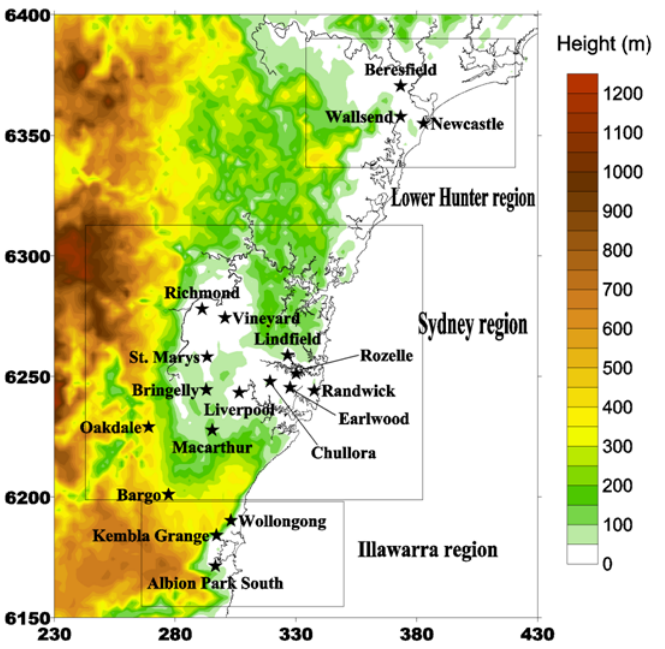

Figure 3: Monitoring sites in the state of New South Wales, Australia which includes Lower Hunter, Sydney and Illawarra region.

Scientific and Industrial Research Organization (CSIRO), in Australia, for use in air quality studies on a local, regional or inter-regional scale [41]. Recently, a modified version of TAPM called TAPM-CTM was developed to include the LCC and carbon bond IV photochemical mechanism as well as the GRS (Generic Reaction Set) photochemical component, which was released in 2008.

As the regulatory agency, the NSW Office of Environment and Heritage is mostly interested in the prediction of peaks ozone scenarios, only episode days are chosen for the simulation in this study. The spatial distributions of 8-hour maximum average of the ozone level are extracted from those simulations for the smallest grid cell (i.e. $2km \times 2km$).

It is noted that there are some differences in the ozone level as predicted by the TAPM-CTM model, compared to the actual measurement data at the monitoring stations. Most of the TAPM-CTM predicted outputs are under-predicted, especially during the episode days. Moreover, their correlation is usually nonlinear, and different from day to day. For correcting the under-prediction and improving the correlation between the model output and the measurement data, the modeled ozone datasets need to be calibrated, e.g. by using the regression analysis via comparison of the actual and the simulated data at all the monitoring stations to determine the correlation ratio between them. For example, Figure 4 shows a correlation of daily 8-hour maximum average of ozone for a day in summer. A regression line is drawn by setting the intercept point at zero. Therein, the correlation ratio is determined as 1.326, i.e. all the daily ozone distribution data from TAPM-CTM output are multiplied with this ratio. This comes from the assumption that the spatial distributions of the pollutant are in general predicted correctly enough with the deterministic model, but it needs further compensation due to the under-predict or over-predict situations. The aim here is to form a dataset that is close to the actual data for the whole domain, based on the available correlation ratio at all monitoring stations, i.e. by a regression technique.

For the NO_x input dataset, the measured concentration data for the same days as the TAPM-CTM simulations are used to compute the variation of the NO_x emission rate. The hourly NO and NO_2 concentration for each day is converted to the emission rate according to their molecular mass values and average wind speeds. The downwind distance is estimated accordingly to cover $2km \times 2km$ grid cells, and the other coefficients are set, based on the environment stability conditions by using the Pasquill Table [39]. The calculated hourly emission rate is summed to get the daily emission rate of NO_x at every monitoring station. The emission values for other cells in the domain are approximated in accordance with the nearest distance to the station at which the wind direction and the cell-station direction make the smallest angle. Within a certain radius from stations, pollutant concentrations are assumed to be similar and hence the same emission rate level is expected. On the other hand, the gridded inventory emission rate data for NO_x are extracted from the TAPM-CTM pre-processing outputs. Finally, both types of emission (i.e. inventory and calculated) for each cell are added to form distributed daily NO_x emissions (in kg/day).

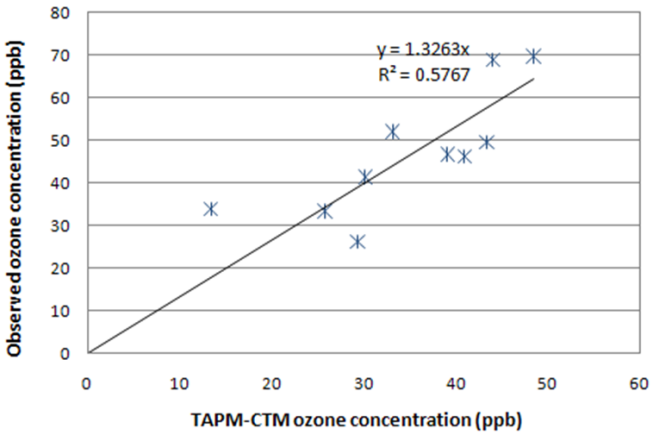

Figure 4: Regression analysis for determination of the correlation ratio between simulated and observed ozone level.

Figure 5 shows a comparison of the daily distribution before and after the summation for a summer day in 2004, where the daily emission is concentrated mostly in the Sydney metropolitan area, especially the area near to the central business district, Sydney East and Sydney inner-West, as shown in Figure 5(a). Obviously, this area has a high population concentration and also dense road networks, as well as a large number of industrial activities. The high emission also appears along the roadways from North to South, and to the West. Figure 5(b) shows that the emission is more scattered in the domain, while it is not distributed well in the East area because there are no measurement data available in that area (the Tasman Sea).

The rest of the input dataset (i.e. coordinate, height from sea level and temperature) can be extracted from the TAPM-CTM model which uses synoptic data collected by the Australian Bureau of Meteorology. The training process is executed by setting the spread parameter at 0.1, the least square weighting coefficient as 1 and the mean square error goal to be 0.004. After several epochs, the network is constructed once the set goal has been met in just 6 minutes of the simulation time. From 2448 patterns of the training dataset (thin size), 343 centres and hidden neurons are used to create the model network.

4.3. Model performance

To validate the trained model, denser datasets (from the same simulation days in the training stage), which involve 21000 data patterns consisting of data collected from January to February 2004, are used. The performance of the validation phase is shown in the scatter plot of Figure 6. It consists of 3500 data points, which correspond to 3500 cells for $2km \times 2km$ each size of the whole domain (i.e. 70 cells to the East 50 cells to the North). The plot represents a correlation between the prediction results by using the constructed RBFNN model against the target outputs in the dataset. As depicted, most of the scatter points are located close to the bisecting line for every data point with the determination coefficient (R^2) of 0.94, which can be considered as good performance.

The spatial distribution, obtained by using the RBFNN model and TAPM-CTM model, is shown in Figure 7. Results of two

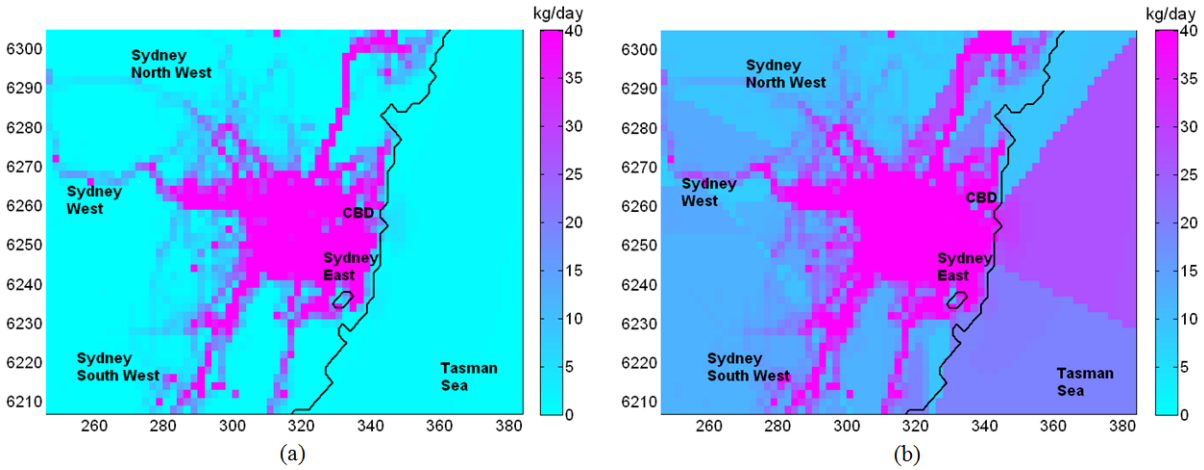

Figure 5: Daily NO_x emission for a day in summer: (a) post-process by TAPM-CTM from the emission inventory, (b) added with the calculated emission.

episode days are presented, wherein both models give similar patterns of the spatial lines but with different ranges of concentrations. For the first day, the higher levels were concentrated at the West area from North to South with range from 21 to 90 *ppb* for the RBFNN, and from 8 to 62 *ppb* for the TAPM model. On the second day, the high concentration scattered about the whole domain in which the peak levels appeared mostly in West area towards South West area. However, RBFNN output gives a maximum level of 103 *ppb* while the maximum level by TAPM is only 72 *ppb*, which exhibits an under-prediction. This uncertainty is confirmed by comparing those levels with actual data collected at the monitoring points.

From these spatial distribution results, it can be observed that most of the high ozone level always appeared, especially during the episode days, in the West of Sydney, including suburban and semi-rural areas. This is the general pattern of ozone occurrence in the Sydney basin which is consistent with meterological conditions of the West and South West being downwind of the sea breeze during the day. In the morning after sunlight, off-shore sea breeze flows from the East and North East across Sydney towards the South West tend to cause an elevated level of ozone in the South West and West of Sydney in the afternoon.

However, the most important issue is the number of exceedance (i.e. more than 80 *ppb* for 8-hour maximum average standard) that are observed, which may have an adverse impact on the human health as well as on the vegetation. This situation rises up due to the increase of the ozone level caused by the accumulation of ozone formed previously in the East of Sydney, which is the transported to the West and South West areas.

4.4. Performance comparison

To assess the reliability of the models, five days simulation results of the spatial distribution are compared with the actual

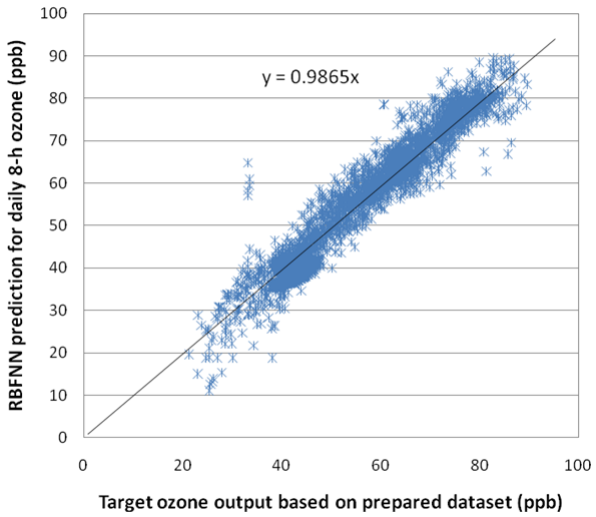

Figure 6: Scatter plot to illustrate the performance of validation phase.

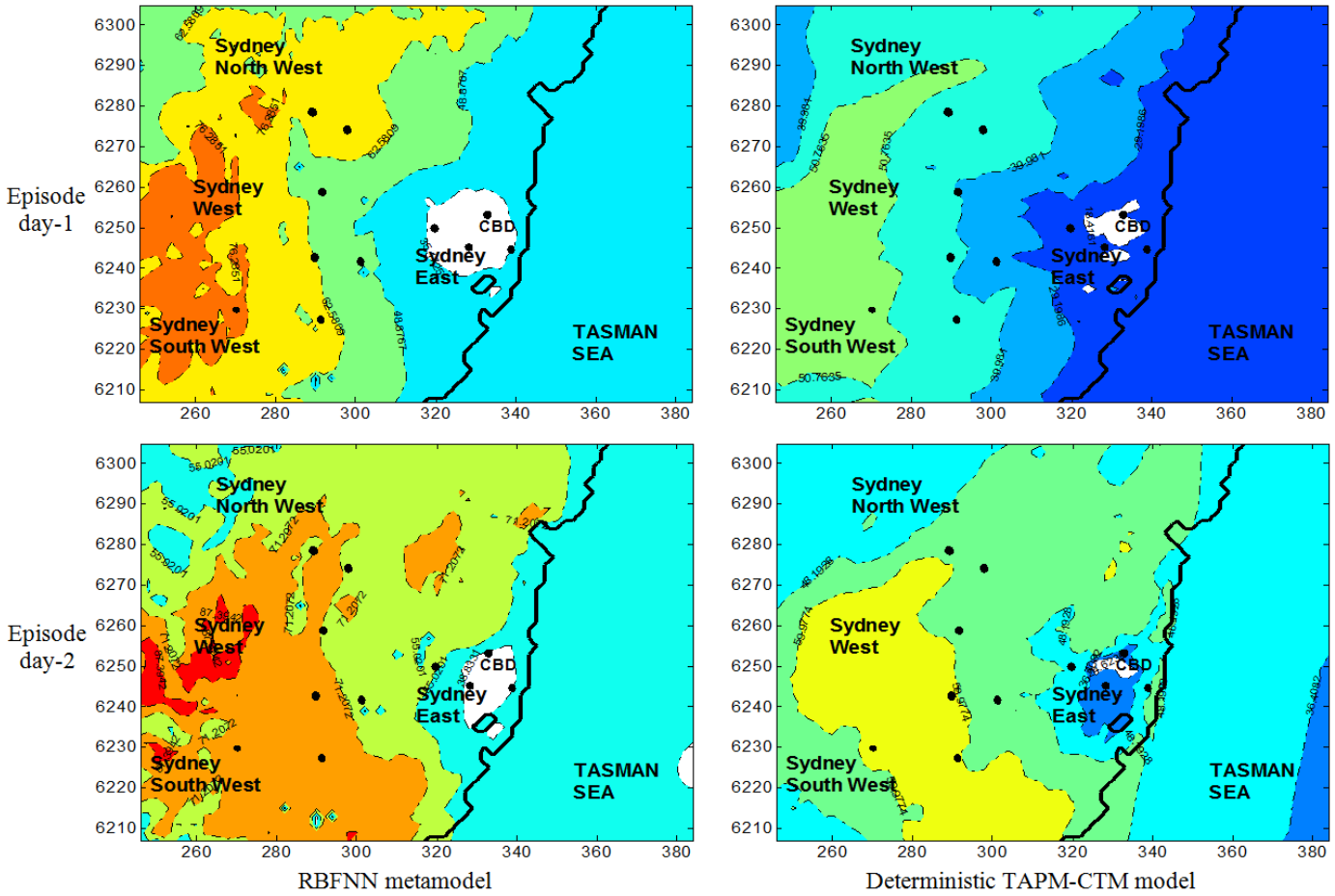

Figure 7: Spatial distribution for 8-hour maximum average of ozone by using RBFNN and TAPM-CTM model (Note: the bullet dots show the location of the monitoring stations).

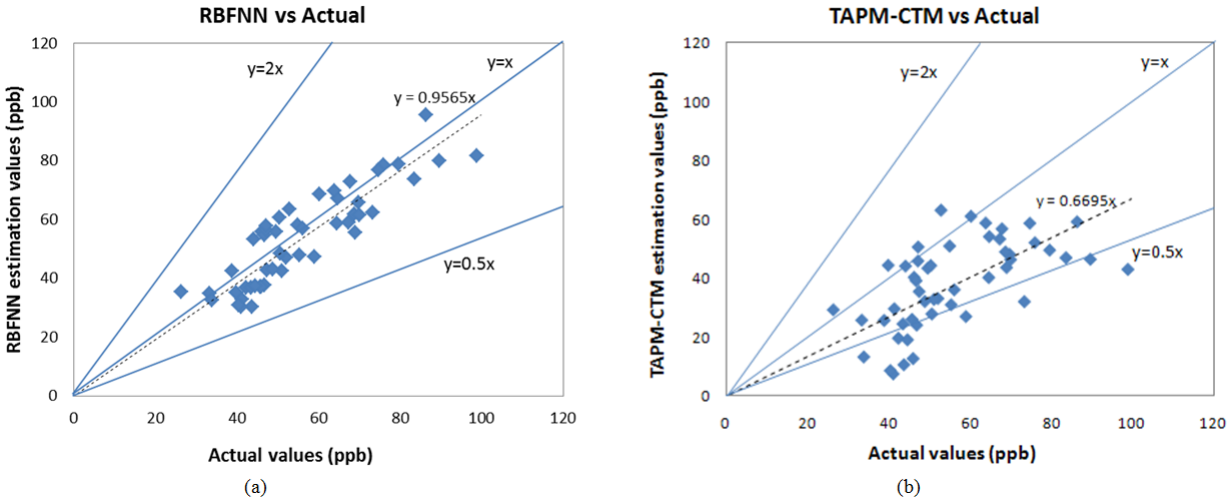

Figure 8: Performance comparison between RBFNN and TAPM-CTM predictions for 8-hour maximum average of ozone at 10 sites in Sydney region.

measurement data at 10 monitoring stations for each day. Figure 8 shows the scatter plots of the models versus the actual data, whereby each plot consists of 50 data points (i.e. 5 days 10 monitoring stations). Five episode ozone days in a summer season are selected in the analysis. As can be seen from the first figure, most data points are located close to the bisecting lines, all lying in between the upper-half section line and lower-half section line. This is an improvement as compared to the TAPM

Table 1: Comparison results for the spatial distribution estimation of pollutant using two RBFNN methods (Note: the *MSE* set to 0.008).

Method	**Spread parameter, σ**	**Weight coefficient, α**	**Performance measure**				**Network size**	**Simulation time (*s*)**
			RMSE	*MAE*	R^2	d_2		
RBFNN-FSWLS	0.1	0.7	12.685	8.460	0.515	0.896	348	228
		0.8	10.202	7.452	0.653	0.913	171	76
		0.9	9.284	7.033	0.713	0.928	96	96
		1.0	9.017	6.894	0.729	0.932	81	33
		1.1	9.103	6.959	0.724	0.931	76	28
		1.2	9.101	6.963	0.726	0.932	74	29
		1.3	9.164	6.973	0.724	0.931	75	29
	0.2	0.7	9.788	7.484	0.681	0.920	167	76
		0.8	9.089	7.134	0.725	0.931	67	25
		0.9	8.798	6.871	0.742	0.936	61	23
		1.0	9.157	6.944	0.721	0.930	35	14
		1.1	9.094	6.968	0.725	0.931	30	12
		1.2	9.089	6.981	0.725	0.931	26	11
		1.3	9.017	7.070	0.729	0.932	36	14
RBFNN-OLS	0.1	-	9.124	7.029	0.723	0.931	90	33
	0.2	-	8.979	6.918	0.731	0.933	31	15

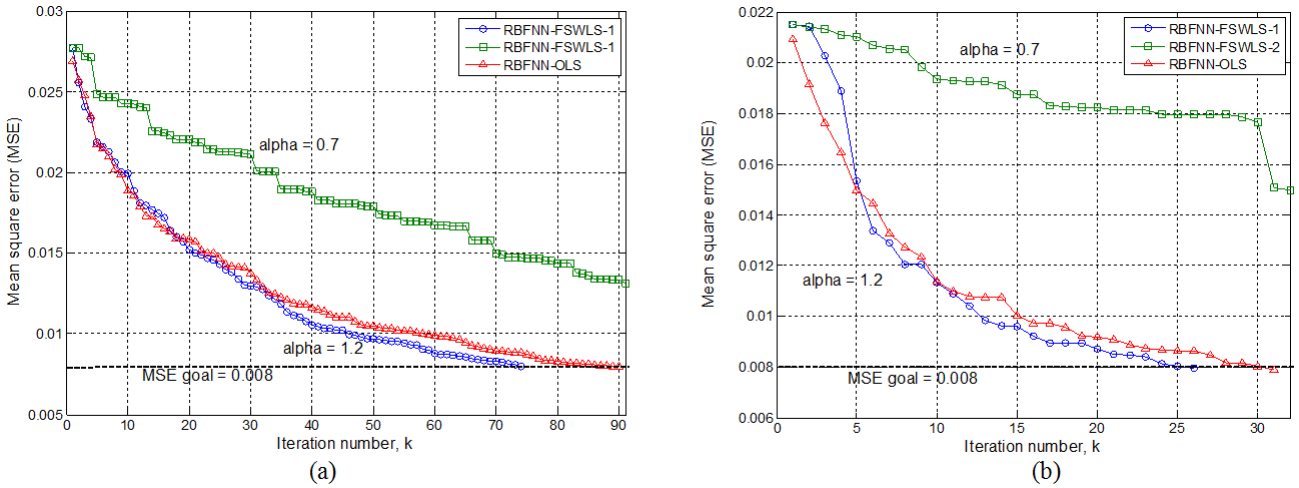

Figure 9: The comparison of the training performance between FSWLS and OLS methods: (a) σ=0.1, (b) σ=0.2.

estimations in which most of the TAPM values show under-prediction results, as presented in Figure 8(b). In terms of R^2 values, RBFNN results in 0.7703 while TAPM gives 0.3521, which can be claimed as another advantage of the proposed approach. However, this indication value shows that further improvements in the approach need to be carried out, as there are some estimation points that do not achieve the actual measurement value. It is probably due to the preparation of the output dataset (for training the model), which much depends on the regression analysis to correlate with the actual measurement data, and on other uncertainty coming from the TAPM-CTM simulation outputs.

In another analysis, the performance of the proposed algorithm for training the RBFNNs centres, featuring the forward selection and the weighted least square (FSWLS), is compared with a typical RBFNN algorithm, i.e. the orthogonal least square (OLS) method [28]. Several values of the spread parameter, σ, and weight coefficient, α, are evaluated, as shown in Table 1. Four performance indexes are used to determine the accuracy of each method as approximation functions, which includes the root mean square error (RMSE), the mean absolute error (MAE) the determination coefficient (R^2), and the index of agreement, d_2. By varying α for FSWLS, it is found that the best performance occurs when α is set to 1.2, for both test values of the spread parameter. Besides, it is learnt that the possible value of α is located between 0.75 and 1.25, to keep the algorithm in the convergence region. As compared to OLS method, at the highest performance by the FSWLS method, the computational cost in terms of the hidden neuron number used and simulation time are slightly improved. The comparison of the training evolution is illustrated in Figure 9. Therein, the OLS method requires 90 and 31 hidden neurons to reach the *MSE* goal of 0.008, while the proposed method only uses 76 and 26 hidden neurons, with the spread parameter between 0.1 and 0.2, respectively.

5. Conclusion

This paper has presented a radial basis function neural network approach to effectively estimate the spatial distribution of daily ozone concentrations with an adequately fast computation time. The model approximates the nonlinear relationship between the NO_x emission, ambient temperature, location coordinates and topography, considered as the inputs, and the 8-hour maximum average of ozone concentration as the output. For the NO_x emission distribution, the emission rate is derived from the measured concentration by using the Gaussian dispersion model, and then added with the emission rate obtained from the emission inventory data. In the training stage, target output data for ozone distribution are extracted from a deterministic air quality model and calibrated to correlate with the actual data obtained from the monitoring stations by using regression. Here, data from the deterministic model and the actual measurements are combined to construct the neural network model to enhance its training performance. Moreover, the proposed approach features the selection RBFNN centres using the forward selection with weighted least squares, offering some performance improvements over the orthogonal least square method to result in a smaller number of hidden neurons used and better estimation results. The methodology is then applied for air pollutant data collected from the monitoring stations in the Sydney basin. The results obtained indicate a promising application of the proposed method in the estimation of ozone concentration with a reasonable accuracy. Compared with the TAPM-CTM model, the proposed method gives higher performance, in which most of the estimated values are closer to the measurement data, while requiring less computation time. The generic methodology indicates that combining a deterministic approach (such as the TAPM-CTM model) and a neural network approach, as proposed in this paper, gives a better estimation of the air pollutant concentration temporally and spatially rather than just using only the dispersion model as currently used by most regulatory agencies.

6. Acknowledgement

This work is supported, in part, by The New South Wales Government through its Environmental Trust, project 2012-RDS-034.

References

[1] A. Elkamel, S. Abdul-Wahab, W. Bouhamra, E. Alper, Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach, Advances in Environmental Research 5(1) (2001) 47-59.

[2] S. Seigneur, Current status of air quality models for particulate matter, Journal of the Air and Waste Management Association 51 (2001) 15081521.

[3] S.B. Phillips, P.L. Finkelstein, Comparison of spatial patterns of pollutant distribution with CMAQ predictions, Atmospheric Environment 40(26) (2006) 4999-5009.

[4] A. Monteiro, A.I. Miranda, C. Borrego, R. Vautard, Air quality assessment for Portugal, Science of The Total Environment, 373(1) (2007) 22-31.

[5] H. Duc, I. Shannon, M. Azzi, Spatial distribution characteristics of some air pollutants in Sydney, Mathematics and Computers in Simulation 54 (2000) 121.

[6] A. Trabelsi, F. Lafont, M. Kamoun, G. Enea, Fuzzy identification of a greenhouse, Applied Soft Computing 7(3) (2007) 1092-1101.

[7] M.H. Fazel Zarandi, M.R. Faraji, M. Karbasian, Interval type-2 fuzzy expert system for prediction of carbon monoxide concentration in megacities, Applied Soft Computing 12(1) (2012) 291-301.

[8] P. Hjek, V. Olej, Ozone prediction on the basis of neural networks, support vector regression and methods with uncertainty, Ecological Informatics 12 (2012) 3142.

[9] B. Ozbay, G.A. Keskin, S.C. Dogruparmak, S. Ayberk, S., 2011. Predicting tropospheric ozone concentrations in different temporal scales by using multilayer perceptron models, Ecological Informatics 6(34) (2011) 242247.

[10] Y. Feng, W. Zhang, D. Sun, L. Zhang, Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification, Atmospheric Environment 45(11) (2011) 19791985.

[11] E.G Ortiz-Garcia, S. Salcedo-Sanz, A.M. Perez-Bellido, J.A. Portilla-Figueras, L. Prieto, Prediction of hourly O_3 concentrations using support vector regression algorithms, Atmospheric Environment 44(35) (2010) 44814488.

[12] M. Boznar, M. Lesjak, P. Mlakar, A neural network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain, Atmospheric Environment. 27(2) (1993) 221-230.

[13] W. Wang, W. Lu, X. Wang, A.Leung, Prediction of maximum daily ozone level using combined neural network and statistical characteristics, Environment International 29(5) (2003) 555-562.

[14] A. Coman, A.Ionescu, Y. Candau, Hourly ozone prediction for a 24-h horizon using neural networks, Environmental Modelling and Software, 23(12) (2008) 1407-1421.

[15] Z. Zainuddin, O. Pauline, Modified wavelet neural network in function approximation and its application in prediction of time-series pollution data, Applied Soft Computing 11(8) (2011) 4866-4874.

[16] A. Pelliccioni, T. Tirabassi, Air pollution model and neural network: An integrated modelling system, Il Nuovo Cimento C 31(3) (2008) 253-273.

[17] C. Carnevale, G. Finzi, E. Pisoni, M. Volta, Neuro-fuzzy and neural network systems for air quality control, Atmospheric Environment 43 (2009) 4811-4821.

[18] H. Pfeiffer, G. Baumbach, L. Sarachaga-Ruiz, S. Kleanthous, O. Poulida, E. Beyaz, Neural modelling of the spatial distribution of air pollutants, Atmospheric Environment, 43 (20) (2009) 3289-3297.

[19] H. Fang, M. Rais-Rohani, Z. Liu, F. Horstemeyer, A comparative study of metamodeling methods for multiobjective crashworthiness optimization, Computers and Structures 83 (2005) 2121-2136.

[20] S.M. Clarke, J.H. Griebsch, T.W. Simpson, Analysis of support vector regression for approximation of complex engineering analyses, ASME Journal of Mechanical Design 127 (2005) 1077-1087.

[21] R.G. Regis, C.A. Shoemaker, Constrained global optimization of expensive black box functions using radial basis functions, Journal of Global optimization 31(1) (2005) 153-171.

[22] L. Ma, K. Xin, S. Liu, Using radial basis function neural networks to calibrate water quality model, World Academy of Science, Engineering and Technology 38 (2008) 385-393.

[23] H. Liu, Y. Wen, F. Luan, Y. Gao, Application of experimental design and radial basis function neural network to the separation and determination of active components in traditional Chinese medicines by capillary electrophoresis, Analytica Chimica Acta 638(1) (2009) 88-93.

[24] N.A. Al-Geelani, M.A.M. Piah, R.Q. Shaddad, Characterization of acoustic signals due to surface discharges on H.V. glass insulators using wavelet radial basis function neural networks, Applied Soft Computing 12(4) (2012) 1239-1246.

[25] D. Broomhead, D. Lowe, Multivariate functional interpolation and adaptive networks, Complex System 2 (1988) 321-355.

[26] Z. Ukyan, C.G. Gzelis, Input-output clustering for determining the centers of radial basis function network, Proc. of ECCTD-98, Budapest, Hungary (1997) 435-439.

[27] H.R. Wang, H.B. Wang, L.X. Wei, Y. Li, A new algorithm of selecting the radial basis function networks center, Proc. Int. Conf. on Machine

Learning and Cybernetics, Beijing, China (2002) 1801-1804.
[28] S. Chen, C. Cowan, P. Grant, Orthogonal least squares learning for radial basis function networks, IEEE Transactions on Neural Networks 2(2) (1991) 302-309.
[29] M.J.L. Orr, Regularised centre recruitment in radial basis function networks, Neural Computation 59 (1993) 1-11.
[30] R. Horn, C. Johnson, Matrix analysis, Cambridge University Press, Cambridge, UK (1985) 18-19.
[31] H. Wahid, Q.P. Ha, H. Duc, M. Azzi, Estimation of background ozone temporal profiles using neural networks, Proc. IEEE Int. Conf. on Intelligent Computing and Intelligent System 3, Guangzhou (2011) 292-297.
[32] H. Demuth, M. Beale, and M. Hagan, Radial Basis Networks, Neural Network ToolboxTM 6, MATLAB Users Guide, (2009) 8-1 8-12.
[33] H. Duc, M. Azzi, H. Wahid, Q.P. Ha, Background ozone level in the Sydney basin: Assessment and trend analysis, Journal of Climatology (2012). DOI: 10.1002/joc.3595
[34] J.H. Seinfeld, Urban air pollution: state of science, Science 243 (1989) 745-752.
[35] R.E. Morris, G. Yarwood, C.A Emery, G.M. Wilson, Recent advances in CAMx air quality model, ENVIRON International Corporation, presented at the A and WMA anual meeting, Orlando, paper No. 934 (2001). Avaialable at http://www.camx.com/publ/.
[36] D. Byun, K.L. Schere, Review of the governing equations, computational algorithm, and other components of the Model-3 Community Multiscale Air Quality (CMAQ) Modelling System, Applied Mechanic Review 59 (2006) 51-77.
[37] U.S. Environmental Protection Agency, Air quality criteria for ozone and related photochemical oxidants (final), Research Triangle Park, NC; EPA/600/R-05/004af (2006).
[38] R. Beier, A. Doppelfeld, Spatial interpolation and representativeness of air quality data: An intuitive approach, Proc. Int. Conf. Air Quality in Europe, Padova Italy (2000).
[39] F. Pasquill, Atmospheric dispersion parameters in Gaussian plume modeling. U.S. Environmental Protection Agency Rep. EPA-600/4-76-030B (1976).
[40] D. Cooper, F.C. Alley, Air Pollution Control: A Design Approach, Waveland Press, Inc., 4th edition, 2011.
[41] P. Hurley, TAPM V4 Part 1: Technical description, CSIRO Atmospheric Research Paper No. 25 (2008).
[42] P.A. Vesilind, J.J. Peirce, R.F. Weiner, Environmental Engineering, Butterworth Heinemann. 3rd ed. (1994).