A Metamodel for Background Ozone Level using Radial Basis Function Neural Networks

Herman Wahid\textsuperscript{1,2} \quad Q. P. Ha\textsuperscript{2} \quad Hiep Nguyen-Duc\textsuperscript{3}

\textsuperscript{1}Faculty of Electrical Engineering 
Universiti Teknologi Malaysia 
Skudai, Johor, Malaysia. 
hbinwahi@eng.uts.edu.au 

\textsuperscript{2}Faculty of Engineering 
University of Technology, Sydney 
Broadway, NSW, Australia. 
quangha@eng.uts.edu.au 

\textsuperscript{3}Department of Environment and Climate Change 
NSW, Australia. 
hiep.duc@environment.nsw.gov.au

\textbf{Abstract}---In air quality modelling, determination of the background ozone level is essential as it highly affects the accuracy of the photochemical air quality model. It is known that the background ozone level, especially in urban areas, has been changing over the years. Unfortunately, the reasons of that alteration were not clear and the background ozone itself was not easily derived in practice. In this paper, a new background ozone model will be developed by using the ozone ambient quality data and the meteorological data at the several stations in the Sydney basin. To accomplish the modelling process, an adaptively-tuned radial basis function neural network metamodel is proposed and utilised in the simulation. Different input parameters are considered to evaluate their influence on the constructed background ozone model. The proposed model, subject to some statistical criteria, demonstrates its capability of estimating the background ozone level with a reasonably good accuracy.

\textbf{Keywords}---metamodelling, radial basis function neural networks, adaptive spread, background ozone

I. INTRODUCTION

The tropospheric ozone, commonly-called surface ozone, contains 10% of the ozone layer residing between the earth surface and the stratosphere layer. It is contributed from several sources, including a portion of the stratospheric ozone that is transported down, chemical reactions of the natural gases and mostly the reaction with the human-made pollutant gases. Unfortunately, excessive surface ozone exposure may be harmful rather than beneficial to the earth’s livings because it could damage living tissues and might contribute to the warming process of the earth surface.

The background ozone level is referred to the level of the ozone occurring at the troposphere, which is naturally-formed and free from anthropogenic influences [1]. Accurately determining the background ozone level, which is essential for air quality modeling and control, requires a clean environment to be free from these anthropogenic influences. Such environment is practically not easy to obtain as its original settings have been changed by humans through anthropogenic processes. A definition for the background ozone by the United States Environmental Protection Agency (US-EPA) is so-called the Policy-Relevant Background (PRB) level [2]. The estimated PRB ozone concentrations are shown to be depending on season, altitude and the total surface ozone. Other authors suggested that the background ozone is dependent on a number of factors such as the nature of upwind flow, pollution sources, and terrain conditions, including deposition with respect to forest or agricultural areas [3]. In this paper, we consider the non-photochemical background ozone level, derived from the ambient measurements during nighttime and early morning hourly values (i.e. from 19:00pm to 08:00am the next morning). It excludes the photochemical processes that would occur during daytime as if only natural sources were present.

The remainder of this paper is organised as follows. Section 2 gives an overview of air quality modelling, the metamodelling methodology as well as radial basis function neural networks. Section 3 describes our neural network-based metamodel for estimation of the background ozone level on a short-term basis, and also the linear regression technique used for trend analysis of the background ozone level. The proposed technique features a learning process that can be made more efficient by adaptively adjusting the spread parameter. Results and discussion are reported in Section 4. Finally, a conclusion is drawn in Section 5.
All the above techniques might provide promising results for long term trend estimation of pollution concentrations, but not for dealing with a short-term basis. To overcome this problem and to reduce the modeling complexity involved in retrieval from meteorological data, the neural network approach offers an efficient way to simplify the mathematical modelling process.

B. Metamodelling

Metamodelling has been a major research field since the last decade whose objectives are to obtain a simpler model from a complex model, to approximate the non-linear system behavior, and to reduce the cost, time, and amount of effort required during a simulation analysis. Some reviews for various metamodelling techniques have been presented, see e.g. [7] and [8]. Substantial results from the existing works illustrate that using metamodels to locate an optimum solution is often sufficiently accurate in many applications requiring prediction, optimisation, verification and validation [9].

The trade-off between accuracy and computational expense as well as between local and global information must be considered when developing a simulation metamodel. Hence, current research into metamodelling has focused on improving its accuracy and computational speed. For example, the support vector regression (SVR) was introduced in [10] for improving the simulation time and accuracy. The radial basis function was extended in [11] to add more flexibility. However, the use of metamodels for air quality simulation models has been very rare to date. In this work, we look into the feasibility of the radial basis function neural network metamodel with the aim to construct a generic model for determining the background ozone level at several monitoring stations in an urban area.

C. Radial Basis Functions

Over the last few years, the radial basis function neural network (RBFNN) has been explored and are being successfully applied across many problem domains that cover engineering, medicine, environment control and geology. It is capable of modelling extremely complex functions with large numbers of input and output variables. Besides, the RBFNN paradigms are global, thus a single neural network could be developed to model the entire simulation response surface. This differs from polynomial regression metamodelling, where the regression surface is fitted to a locality, i.e. a subset of the response surface [12].

The RBFNN is motivated by the locally tuned response observed in biological neurons. The main architecture of the RBFNN consists of two layers: (i) a single hidden layer and (ii) an output layer. The nodes in the hidden layer are associated with centres, with a dimension equal to the number of inputs. The output of the RBFNN is mathematically represented as follows:

\[ y_i = f(x) = \sum_{k=1}^{N} w_k \phi(\|x-c_k\|) i = 1,2,3,...m \]  

(1)

where \( i \) is the output index, \( m \) is the maximum output number, \( x \in \mathbb{R}^{n_0} \) is an input vector, \( \phi() \) is a basis function, \( \| \cdot \| \) denotes the Euclidean norm, \( w_k \) are the weights in the output layer, \( N \) is the number of neurons (and centres) in the hidden layer and \( c_k \in \mathbb{R}^{n_0} \) are the RBF centres in the input vector space.

Some of the most commonly used basis functions include Gaussian, multiquadric, inverse multiquadric and thin-plate spline. In this work, Gaussian functions will be used, given by

\[ \phi(\|x-c\|) = \exp\left(\frac{\|x-c\|^2}{2\sigma^2}\right) \]  

(2)

where \( \sigma \) is the standard deviation (or spread parameter) and \( c \) is the centre vector of the Gaussian function.

III. METHODOLOGY

A. RBFNN with adaptively tuned spread

There are two difficulties involved in designing a RBFNN: first to obtain the location of centres for radial basis function, and second, the initial configuration of an RBFNN to be determined by trial and error. Several approaches have been presented to solve the first problem, which include orthogonal least squares [13,14,15], genetic algorithm [16,17], and gradient decent [18].

The selection of a spread parameter (\( sp \)) is also crucial as it will affect the variance of width to the radial basis function. The significance of considering this point in order to obtain the best network has been remarked in [19] and [20]. For example, as mentioned in [19], it was suggested that \( sp \) must be greater than 0.1 of the interval between inputs, and less than 2 of the distance between the leftmost and rightmost inputs. But, it is still unclear about an optimal value for \( sp \).

By experimenting with a test function, it is noted that the network error was decaying to a minimum point while varying the spread parameter value, in which the minimum point was considered as the optimal point of the spread parameter [21]. The optimal value is different each time the new neuron is added to the hidden layer of the network. This illustration prompts to the suggestion that an optimal value for the spread parameter is a function of RMSE at the minimum global point. Consequently, we could use the gradient criteria to solve this problem, where the minimum point can be obtained when the first derivative of RMSE with respect to \( sp \) approaches 0, i.e.

\[ \text{gradient } s = \nabla s = \frac{\partial (\text{RMSE})}{\partial sp} = 0. \]  

(3)

By giving an initial value of a spread parameter (\( sp_0 \)), the optimal point can be solved by using an optimisation technique. It has been determined by many experiments that the \( sp_0 \) must be chosen between 0.1 and 1 in order to get the best convergence during the optimisation process. In this work, we use the steepest descent [22] technique for simplicity, given that this method may appear to the best unconstrained minimisation although it may yield a local minimum rather than the global minimum. The procedure can be written as

\[ sp_{(new)} = sp_{(old)} - \nabla s, \]  

(4)
where $\gamma > 0$ is the step size (or learning rate) and $\nabla_S$ is the gradient derived in (3). To end the iteration process, a termination criteria was used, where the change in the function value for two consecutive iterations becomes smaller than a given threshold.

Furthermore, a pruning algorithm is also considered during the adaptation of the spread, where the number of basis functions in the network can be either increased or decreased over-time to avoid over fitting or under fitting. The steps begin with computing the mean value of each radial basis output in the hidden layer, as expressed in the following equation:

$$\bar{\phi}_i = \frac{1}{n} \sum_{j=1}^{n} p_j, \quad i = 1, 2, 3...m, \quad (5)$$

where $\phi$ is the hidden layer output, $i$ is the neuron number, $j$ is the input number, and $p$ is the input vector. It is followed by normalising those outputs between 0 and 1, which results in a series of normalised $\phi$ output, as given in the following equation,

$$D_i = \left| \frac{\phi_i}{\phi_{i\text{ (max)}}} \right|, \quad i = 1, 2, 3...m. \quad (6)$$

When the $D_i$ value is less than a certain threshold, $\delta$ during the training process, the $i^{th}$ hidden node could be removed.

B. Background ozone level estimation model

The idea in our metamodelling approach is that we can design a network model for each monitoring station, which, from meteorological data collected, we could construct a more generic model that may extend over the whole area in the Sydney basin. In this study, our scope is to model a network to predict the background ozone level at several urban sites in Sydney, namely Randwick, Blacktown and Vineyard. A proper selection of the input and output characteristic is essential in order to make the RBFNN learn with a fast convergence. Typically, more input data are better to make the model more comprehensive and the interpretation more convincing. In this paper, we analysed several input variables including pollutant data and the related meteorological data, to look into their variation effect on the accuracy of the constructed model. The pollutant data include the concentrations of ozone ($O_3$), nitrogen oxide (NO), nitrogen dioxide ($NO_2$), carbon monoxide (CO), sulfur dioxide ($SO_2$), volatile organic compound (VOC), particulate matters of size $<10\mu m$ (PM$_{10}$), and size $<25\mu m$ (PM$_{2.5}$). Other meteorological data include the wind speed (WSP), wind direction (WDR), and the air temperature (TEMP). In addition, the time information was also considered as an input variable.

The background ozone level has been set as the target output of the network. From the earlier definition, the non-photochemical background ozone level is considered. For further elaboration, it is defined as the ozone concentration where the NOX agent is set to be zero (i.e. NOX = NO + NO$_2$). Fiore et al. suggested that the estimation could be more accurate if the correlations with reactive nitrogen oxides ($NO_X$) are taken into consideration [23]. It implies that a component of the background ozone is produced from natural precursor sources, which include NOX emission from soil and lightning, and hydrocarbon emissions from vegetation.

We used the hourly data collected by the Department of Environment, Climate Change and Water (DECCW), New South Wales. These data have been carefully recorded at various stations in the Sydney basin and post-processed to be trusted as good data sets. The data used in this study cover a period of 5 months from November 2000 to March 2001. However, it is difficult to get the condition when the NOX is equal to zero at every hour especially at the daytime, as the photochemical effects were dominant during day time. Therefore, we only considered for the evening and the early morning ozone data (i.e. from 19:00pm to 08:00am the next morning) for the background ozone level, in which the anthropogenic effect can be minimised. Unfortunately, there still exist a number of hourly data whose background ozone cannot be determined. Thus we replaced those missing data by the linear regression of the previous and subsequent measured values at the station. This can be done by taking the local correlation of the ozone concentration and the NOX data, $O_3$:NOX. The intercept of the linear regression line from this correlation indicates the background ozone level at the particular missing point, which could be approximated as the ozone concentration when NOX is equal to zero. Fig. 1 shows an example of the $O_3$:NOX correlation at an absence point, in which we take the data at two hours before and two hours after for that regression. This procedure was repeated for every missing point for the entire data. The intercept value is commonly given by the following equation,

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2}, \quad (7)$$

where in this case, $b$ is the estimated background ozone level (in ppb) at the missing point, $y$ is the ozone concentration (in ppb), $x$ is the NOX concentration (in ppb), and $n$ is the number of data used for the linear regression.

![Fig. 1. Correlation of $O_3$ and NOX to estimate the background ozone level at an absence point.](image)
Furthermore, we prepared two groups of data, which were for the training and testing purposes. For the training, the input and target data were taken from the first and second month, whereas the remaining month’s data can be used as the testing data sets for validation. Besides, the entire wind direction data (WDR) are converted from degree to radian unit, i.e. \( WDR = \frac{2\pi \theta}{360} \). The entire inputs and targets have to be normalised (e.g. in the interval between 0 and 1), e.g. by using \( \text{mapminmax} \) function in MATLAB, in order to make them contribute with the same influence to the RBFNN. After the learning process is finished and the accuracy obtained with some test sets is satisfactory, the network can be used for the prediction with other available data.

C. Performance indices

In the metamodelling approach, it exists three distinct components: experimental design, the metamodel type, and the validation method [24]. The experimental design involves how points are selected in order to properly capture the desired properties. The metamodel type is about how to create an approximation, which in this study, we used RBFNN. The validation method is how the accuracy of the metamodel is evaluated in order to ensure that the metamodel reflect the actual model. Hence, some statistical criteria will be used to measure the residual errors, including the root mean square error (RMSE), the mean absolute error (MAE) and the determination coefficient \( (R^2) \), given respectively as

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (O_i - P_i)^2}{n}},
\]

\[
MAE = \frac{\sum_{i=1}^{n} |P_i - O_i|}{n},
\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (O_i - \overline{O})^2}{\sum_{i=1}^{n} (O_i - P_i)^2},
\]

where \( P_i \) and \( O_i \) are the predicted and observed concentrations, and \( \overline{O} \) represents the observation mean.

In addition, we also investigate the index of agreement, \( d^2 \), a measure expressing the degree to which predictions are error-free [25]:

\[
d^2 = 1 - \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (|P_i - \overline{O}_i| + |O_i - \overline{O}|)^2}.
\]

IV. RESULTS AND DISCUSSION

A. Case A: Randwick station

The Randwick site is located at the east region of Sydney. As reported in [2], this urban site was recorded as having the highest average of non-photochemical background ozone level in the east region from the years 1998 to 2005.

In this study, the analysis has been done using the data from November 2000 to March 2001. We performed five simulations of different combinations of inputs as shown in Table 1 using the proposed algorithm. The first simulation was performed using the time information, the photochemical concentration and the ozone measurements. For the next three sets, we added the first set with meteorological information that includes the air temperature, the wind speed and the wind direction with three combinations. For the entire simulations, we initialize the training process by setting the initial spread parameter to be 0.1 and error goal to be 0.005. After several epochs that depend on the inputs data used, the network is constructed once the set goal has been met.

A good compromise between the network size and the selected variable inputs in the background ozone estimation shows that the best results in terms of coefficient and index of agreement are obtained by using the inputs in the set number 3 as depicted in Table 1. The network size is 17, which means that 17 radial basis functions exist in the hidden layer of the network. The network size is increasing gradually with the number of input variables, according to the growing complexity of the training process. Furthermore, it can be learnt that the wind speed and wind direction affect much the accuracy of the constructed model. We expect that the temperature will give an influence on the performance, unfortunately the results do not support that as shown in the set number 4. It is probably due to a small variation of the recorded temperature level, hence minorly affecting the convergence.

An example of approximation on the testing set from the best results is shown in Fig. 2. As we can observe from the figure, most of the values have shown good results of prediction, in which the predicted values follow the pattern of the actual values derived from ambient measurements. As observed from the graph, the predicted values vary from 6.5ppb to 47ppb, where it demonstrated an increasing trend in the three months with the rate of about 0.002ppb/hour using the linear fit line.

B. Case B: Blacktown station

Blacktown, which is an urban site located at the Sydney West region, was investigated. This station was also showing the highest average background ozone level in this region from the years 1998 to 2004 [2].

The same procedure explained previously for Randwick site has also been applied for the Blacktown station. Five combination sets of input variables were used, with the additional carbon monoxide as the pollutant agent. During the training process, we set the initial spread to be 0.1 and the error goal to be 0.006.

By referring to Table 2, the best combination with the highest accuracy is given by the inputs in the set number 4. Again, meteorological parameters are determined to be dominant in influencing the model performance. Unfortunately, the produced errors as derived by the RMSE and MAE are slightly higher than was generated by the Randwick’s model. Furthermore, the correlation given as a
determination coefficient $R^2$ was also low. The deviation performance may be due to only a small number of non-photochemical conditions were obtained from the dataset while most of the background ozone level data have been determined by using the linear regression method. An example of prediction results is depicted in Fig. 3. The predicted background ozone level data were found to be between 4ppb to a maximum of 40ppb, with a decreasing trend of 0.003ppb/hour.

Table I. Performance Indices on the Test Set for the Simulation Performed by Different Combination of Inputs (at Randwick).

<table>
<thead>
<tr>
<th>Set</th>
<th>Inputs</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>$d_2$</th>
<th>Network size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Time, NO, NO₂, O₃</td>
<td>4.878</td>
<td>3.560</td>
<td>0.464</td>
<td>0.866</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>Time, NO, NO₂, O₃, Temp</td>
<td>5.244</td>
<td>3.949</td>
<td>0.381</td>
<td>0.845</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>Time, NO, NO₂, O₃, WSP, WDR</td>
<td>4.850</td>
<td>3.635</td>
<td>0.471</td>
<td>0.868</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>Time, NO, NO₂, O₃, Temp, WSP, WDR</td>
<td>6.069</td>
<td>4.859</td>
<td>0.171</td>
<td>0.793</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>Time, NO, NO₂, O₃, Temp, WSP, WDR, PM₁₀, SO₂</td>
<td>5.677</td>
<td>4.444</td>
<td>0.275</td>
<td>0.819</td>
<td>27</td>
</tr>
</tbody>
</table>

Table II. Performance Indices on the Test Set for the Simulation Performed by Different Combination of Inputs (at Blacktown).

<table>
<thead>
<tr>
<th>Set</th>
<th>Inputs</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>$d_2$</th>
<th>Network size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Time, NO, NO₂, O₃</td>
<td>11.241</td>
<td>7.483</td>
<td>0.223</td>
<td>0.806</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Time, NO, NO₂, O₃, Temp</td>
<td>11.163</td>
<td>7.365</td>
<td>0.234</td>
<td>0.808</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Time, NO, NO₂, O₃, WSP, WDR</td>
<td>11.590</td>
<td>7.917</td>
<td>0.174</td>
<td>0.793</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>Time, NO, NO₂, O₃, Temp, WSP, WDR</td>
<td>11.025</td>
<td>7.479</td>
<td>0.252</td>
<td>0.813</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>Time, NO, NO₂, O₃, Temp, WSP, WDR, PM₁₀, SO₂</td>
<td>11.727</td>
<td>7.876</td>
<td>0.154</td>
<td>0.789</td>
<td>32</td>
</tr>
</tbody>
</table>

Table III. Performance Indices on the Test Set for the Simulation Performed by Different Combination of Inputs (at Vineyard).

<table>
<thead>
<tr>
<th>Set</th>
<th>Inputs</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>$d_2$</th>
<th>Network size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Time, NO, NO₂, O₃</td>
<td>12.978</td>
<td>8.495</td>
<td>0.283</td>
<td>0.821</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>Time, NO, NO₂, O₃, Temp</td>
<td>12.447</td>
<td>8.336</td>
<td>0.340</td>
<td>0.835</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>Time, NO, NO₂, O₃, WSP, WDR</td>
<td>12.915</td>
<td>8.750</td>
<td>0.290</td>
<td>0.822</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>Time, NO, NO₂, O₃, Temp, WSP, WDR</td>
<td>12.501</td>
<td>8.610</td>
<td>0.334</td>
<td>0.834</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>Time, NO, NO₂, O₃, Temp, WSP, WDR, PM₁₀, SO₂</td>
<td>13.373</td>
<td>9.198</td>
<td>0.238</td>
<td>0.810</td>
<td>42</td>
</tr>
</tbody>
</table>

C. Case C: Vineyard station

Vineyard is considered as a semi-rural region located at North West of Sydney. Even though, the recorded data from a previous report [2] show that the average background ozone level for six years from 1998 was quite high that could be equivalent to an urban site. It is one proof to show that the high ozone level could also be experienced at a rural site which has a less photochemical effect because it might be transported from a long distance by wind.

For the training process, we use the same procedure and the same input combination that has been done at Randwick. The error goal was set to be 0.0045 with the initial spread of 0.1. From Table 3 we can observe that the best performance based on the index of agreement is given by the combination of inputs as shown in the set number 4. Notably, the network size is smaller than the size used in the set number 3 with less input data. The overall performance is not as good as that of the Randwick’s model. The reason may be from the absence of much non-photochemical data, hence the determination of the background ozone level is much depending on the linear approximation.

Fig. 4 shows the prediction results at Vineyard station from sampled data in the year of 2001. The prediction follows the background ozone pattern of the actual data obtained from measurements at most of the hourly data. However, some predictions cannot be obtained accurately especially when the observed data are peaking. The predicted values are varying from the minimum of 2ppb to a maximum of 22ppb, which shows a decreasing trend with a rate of 0.004ppb/hour.

V. Conclusion

A new metamodel using adaptive radial basis function neural networks with a pruning algorithm have been presented for predicting the background ozone level in several monitoring stations in the Sydney area. The results obtained indicate the promising application of the proposed method in the analysis of biogenic trends and emission impact assessment for air quality modelling. Several combinations of the input variables have been used to evaluate their influence to the accuracy of the constructed model. The considered inputs include time, photochemical precursor pollutants, meteorological data, and other particulate matter data. It has been observed that the photochemical data with the existing of meteorological data, especially the wind speed and wind direction dominantly affect the model performance. It is noted that the model performance does not depend much on the existence of other pollutant agents such as SO₂, CO and PM₁₀. Their influence is probably not crucial for the background ozone prediction, hence their existence in the training data may cause some reduction of the model accuracy. For future work, we would extend the current methodology to be used as the long-term prediction of the background ozone level.

REFERENCES

Fig. 2. Predicted background ozone level and the observed data at Randwick station over night-time and early morning hourly data.

Fig. 3. Predicted background ozone level and the observed data at Blacktown station over night-time and early morning hourly data.

Fig. 4. Predicted background ozone level and the observed data at Vineyard station over night-time and early morning hourly data.


