New sampling scheme for neural network-based meta-modelling with application to air pollutant estimation

H. Wahid [1,2]*, Q.P. Ha [2], and H. Duc [3]

[1] *Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor, Malaysia*
[2] *Faculty of Engineering and IT, University of Technology Sydney, NSW, Australia*
[3] *Office of Environment and Heritage, NSW, Australia*
** Corresponding author (hbinwahi@eng.uts.edu.au)*

Purpose A new method for the design of experiments (DOE) or sampling technique is proposed, using a distance weight function and the *k*-means theory. The radial basis function neural network metamodelling approach[1] is used to evaluate the performance of the proposed DOE by using an *n*-degree of test function, applied to the complex nonlinear problem of spatial distribution of air pollutants. A comparison study is included to analyse the performance of the proposed technique against available methods such as the *n*-level full fractional design method and the Latin Hypercube Design method. **Method** For one design objective and *n* number of input design variables, a set of input-output training dataset are $X = \left\{x_1^{(1)}, x_1^{(2)}, ..., x_1^{(i)}; ...; x_j^{(1)}, x_j^{(2)}, .., x_j^{(i)}) \middle| \, i = 1,.., m, \; j = 1,.., n \right\}$ and $Y = \left\{ y^{(1)}, y^{(2)}, ..., y^{(i)} \middle| \, i = 1,2,..., m \right\}$, where *m* is the maximum number of the data points. Each data point has its own unique weight obtained from the distance factors between point p^i and a common reference point *c*, by using the Euclidean distance measure (i.e. $d_i(p^i, c)$). The weights represent the distinct patterns between each data point. A neighbour can be clustered as a group where the data point is taken as a candidate. To generalise the solution, the pairs of the input and output data points are combined to become the design space, given as $S = \{X; Y\}$.The solution can be simplified further if we set a common reference centre at the coordinate origin by firstly normalising the design space to $\hat{S} = [-1, 1]^{n+1}$. A list of distance weight values, $D = \{d_1, d_2, ...d_i \mid i = 1,2,..., m\}$, is then sorted and clustered by using an available clustering algorithm. In this work, the *k*-means algorithm based on the Voronoi iteration[2] is used due to its fast computation especially in the 1-dimensional case. Here, the initial points are replicated randomly, to expectedly result in a global minimum solution. The maximum number of *k* corresponds to the number data points that will be sampled. **Results & Discussion** To initially validate the accuracy of the scheme, a known test function called as "Hock–Schittkowski Problem 100" is used in which this nonlinear problem involving of 7 variables, 1 objective, and 4 constraints. A prepared dataset which generated randomly, are sampled at different sample size *N*, and then mapped using RBFNN metamodel.

***Keywords**: Automation, sampling, experimental design, metamodel, radial basis function network, air pollution model*

INTRODUCTION

Nowadays, simulation modelling becomes a popular tool for the analysis of complex systems' behaviour. Its popularity is due to the flexibility in the implementation and the ability to model a real world physical system to a certain detail. A deterministic type model has shown good performance in its estimation, but encounters difficulties in the development and highly computational cost in the execution. Thus, metamodels (or surrogate models) have been suggested to be an approximate model that can adequately represent the intrinsically non-linear and complex relationship between the system's input and output.

Before executing the function approximation in metamodelling, it is important to select the design points in the domain which is generally termed as sampling, experimental design, or design of experiment (DOE). The aim of any sampling method is to effectively cover the design space and to gather the essential information of the design space characteristics. These sets of independent design variable values from the data points are utilised to produce the values of dependent variables (i.e. responses), in a process known as computer experiments. Various sampling approaches appeared in the literature such as the full factorial design technique, stratified random sampling, Latin square sampling and Latin Hypercube sampling[1].

This paper presents a new strategy for a metamodel DOE, based on the distance measure and clustering process, referred to as the weighted clustering design (WCD) throughout this paper. Here we employ

the proposed sampling method to develop a radial basis function neural network (RBFNN) metamodel as a function approximator. We first test the scheme validity with a known nonlinear function. The air pollutant estimation problem is then tackled by using the improved metamodel.

The rest of this paper is arranged as follows. An overview on experimental design is discussed after the introduction. The following section describes in detail the methodology which covers the proposed sampling scheme, the implementation of RBFNN as a metamodel, and the performance measure to validate the model. The next section presents the results together with a discussion on the underlying simulation problem, followed by some concluding remarks given in the final section.

OVERVIEW OF SAMPLING SCHEME

The primitive experimental design involves the selection of few data points located at the bounds of the design space, and is called the full factorial array. This is a physical trial method in which the effectiveness of using these points remains very poor. In the computers' era, the experimentation became less costly and the space filling experimental designs started to be used. The full factorial design (FFD) is the simplest sampling approach which is the most general and standard DOE used over the years for the function approximation purpose[1,2]. In FFD, the bounds of all the design variables are firstly identified and then discretised into equal intervals within the design space. For example, for *n*-level FFD, the total *n* points selected for each design variable *v* are equally spaced over the range. It means the number of design points will be n^v. This approach is also known as the rectangular grid point sampling.

Another space-filling method called the Plain Monte Carlo sampling involves using a random number generator to select the points to reduce the number of points in a trial set. While being computationally-efficient, Monte Carlo sampling provides no robustness in finding a space filling set of points[3].

The Latin Hypercube Design (LHD), proposed by McKay et al.[4], is a more sophisticated sampling scheme and continuously being researched. Instead of using all equally-spaced points in the allowable design space, these points are effectively scattered, spanning the whole domain. For selection of *n* number of sample points, the range of each design variable is divided into the same number of non-overlapping regions based on the type of probability distribution function (PDF) specified, which can be either normal or uniform PDF. One segment is chosen from each region at random to form each trial point. As there is no guarantee for a balanced set to be obtained from the points, many researchers have extended the McKay method into optimal LHD[5], inherited LHD[6], and hybrid LHD[7].

METHODOLOGY

Metamodelling research has been a major research field during the last decade[8,9]. Basically, metamodels are constructed in three stages, i.e. preparing the data and choosing the modelling approach; parameter estimation and training; and model validation[1] and testing. In neural network based metamodelling, the data sampling is necessary to reduce the computational burden especially when dealing with a large dataset. The methodology of the proposed sampling scheme is described in the following subsection.

The proposed sampling scheme

Here, a dataset are normally divided into two, one for the training (trial) and another one for the testing. If we have a set of input-output training dataset denoted by *x* and *y*, a mapping solution is given as follows:

$$\left\{x^{(i)} \rightarrow y^{(i)} = f(x^{(i)}) \middle| \; i = 1,2,...,m\right\}, \qquad (1)$$

where *m* is the maximum number of the data points. For the case of one design objective and *n* number of input design variables, the input and output are given as in the following equations,

$$X = \left\{x_1^{(1)}, x_1^{(2)}, ..., x_1^{(i)}; ...; x_j^{(1)}, x_j^{(2)}, .., x_j^{(i)}) \middle| \; i = 1,..,m, \; j = 1,..,n\right\}$$

and

$$Y = \left\{y^{(1)}, y^{(2)}, ..., y^{(i)} \middle| \; i = 1,2,...,m\right\}. \qquad (2)$$

For a 3-dimensional problem, the distribution of four data points is illustrated in Fig.1. Each data point has its own unique weight by measuring the distance weight factors from a common reference point *c*. By using the Euclidean distance measure, the distance between point p^1 and c is mathematically written as,

$$d(p^1, c) = \left[\left(p^1(x_1) - c(x_1)\right)^2 + \left(p^1(x_2) - c(x_2)\right)^2 + \left(p^1(x_3) - c(x_3)\right)^2\right]^{1/2}, \qquad (3)$$

or generally, the weight for all data points of the *n*-dimensional problem is given as follows:

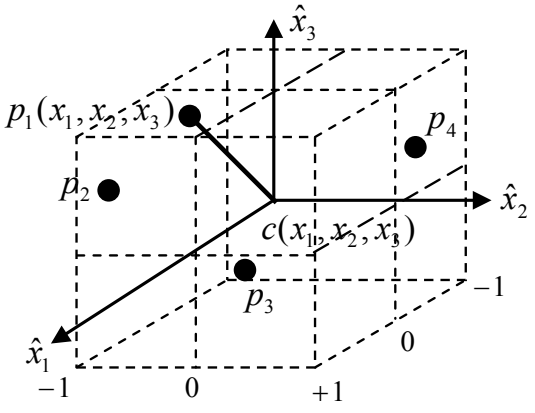

Fig.1. A distribution of data points in 3-dimensional space.

$$d(p^i, c) = \left[\sum_{j=1}^{n} \left(p^i(x_j) - c(x_j) \right)^2 \right]^{1/2} . \quad (4)$$

The weights could represent the distinct patterns between each data point, and some neighbour points may have about similar weight that could be clustered as a group and one point is taken as a candidate.

To generalise the solution, the pairs of the input and output data points are combined to become the design space *S* in this evaluation, which is given as

$$S = \{X; Y\}. \quad (5)$$

Hence, the dimension of distance measures for one targeted output now becomes (*n+1)×m*. The solution (4) can be simplified further if we set a common reference centre at the origin 0 by firstly normalising the design space *S* to the minimum of -1 and to the maximum of 1, i.e.

$$S' = [-1, 1]^{(n+1)\times m}, \quad (6)$$

as shown in Fig.1. Thus, solution (4) now becomes

$$d(p^i) = \left[\sum_{j=1}^{n+1} \left(p^i(\hat{x}_j) \right)^2 \right]^{1/2} , \quad (7)$$

where $\hat{x}_j$ is the normalised values of the design space *S* which has been incorporated by the output variable.

A list of distance weight values

$$D = \{d_1, d_2, ...d_i \mid i = 1,2,..., m\} \quad (8)$$

is then sorted and clustered by using available clustering algorithm. In this work, a well-known *k*-means algorithm based on Voronoi iterations[10] is used due to its fast computation especially for the 1-dimension case. It uses a two-phase iterative algorithm to minimise the sum of point-to-centroid distances, summed over all *k* clusters. There are several methods to choose the initial *k*-means points. In this evaluation, we replicate them randomly, which typically results in a solution that is a global minimum[11]. The maximum number of cluster *k* corresponds to the number data points to be sampled. The determination of an appropriate *k* value for this scheme is demonstrated in this work.

Radial basis function network as metamodel

There are a number of metamodelling techniques that have been researched such as polynomial regression, splines, neural networks, Kriging and support vector machine[8,12]. A few attempts have been made to employ radial basis function neural networks as the metamodelling technique (see e.g. by Liu et al.[13] and Ma et al.[14]).

The RBFNN is a special type of feed forward neural network architecture which consists of an input layer, a hidden layer and an output layer. The neurons in the hidden layer work are the processing elements to perform a non-linear transformation of the input data to approximate the output data. The RBFNN's *q* outputs corresponding to the input vector $x \in \Re^l$ is mathematically represented as follows:

$$f_i(x) = \sum_{k=1}^{n} w_{ki} \phi\left(\| x - c_k \|_2 \right), \quad i = 1,2,3....q \ , \quad (9)$$

where $\phi(\cdot)$ is a basis function, $\|\cdot\|_2$ denotes the Euclidean norm, w_{ki} are the weights in the output layer, *n* is the number of neurons (and centres) in the hidden layer and $c_k \in \Re^n$ are the RBF centres in the input vector space.

In practice, several forms of the basis function ϕ are used for RBF models, and Gaussian is probably the most popular one because it has attractive mathematical properties of universal and best approximation, and its bell shape is easy to control with the spread parameter σ. In this work, a generalised radial basis function neural network (RBFNN) approach is considered, in which fixed biases b are added at the outputs to solve the ill-posed problem relating to singularity (e.g. an approach to regularise the network)[15]. For a Gaussian GRBFNN, equation (9) becomes

$$f_i(x) = \sum_{k=1}^{n} w_{ki} \exp\left(- \| x - c_k \|_2^{\,2} / 2\sigma^2 \right) + b \ . \quad (10)$$

At the hidden and output layer, the position of the radial basis centres, the variance (spread) and the associated linear weights are all unknown parameters that have to be updated. A supervised learning process using a forward selection procedure[16] is implemented to select the position of the centres and a linear least square method is used to train the weights and biases of the output layer.

Metamodel validation

The accuracy of the estimation via metamodelling is evaluated to ensure that the metamodel reflects the actual model. In this evaluation, some statistical indexes will be used for the residual errors, including the root mean square error (*RMSE*), the mean absolute error (*MAE*) and the determination coefficient (R^2), and to use the index of agreement, d_2, a measure expressing the degree to which predictions are error-free[17].

TEST RESULTS AND ANALYSIS

To evaluate the effectiveness of the proposed approach, we first use a benchmark test problem, namely, the Problem 100 from Hock–Schittkowski[18], and then apply it to the problem of estimating the spatial distribution of the ozone concentration in air pollution modelling. For each problem considered, different sample sizes and fitting design methods including weighted clustering design (WCD), n-level full factorial design (n-FFD) and Latin Hypercube design (LHD), are considered. Some of the perform-

ance measures, the size of RBFNN metamodel and the total execution time for the simulations are recorded.

Benchmark Test: Multidimensional function

The Hock–Schittkowski Problem 100 is a test problem consisting of seven variables, one objective, and four constraints[19]. In this analysis, we consider only the objective function without constraints. The design domain for this function is given by

$$f(x_i) = (x_1 - 10)^2 + 5(x_2 - 12)^2 + x_3^{\,4} + 3(x_4 - 11)^2 + 10x_5^{\,6} + 7x_6^{\,2} + x_7^{\,4} - 4x_6x_7 - 10x_6 - 8x_7, \quad (11)$$

where $-10 \le x_i \le 10$. To prepare a full large dataset, a series of input-output data points are randomly generated (e.g. using *'randn'* code in Matlab) within the design space in which the maximum number of data points is set to 4000.

By using three experimental design methods, the prepared data are sampled at a different sample size, *N*. Each set of the sampled data is then mapped using the RBFNN metamodel by setting the spread parameter as 4 and the prescribed *mse* goal as 0.001 for all the testings.

Table 1 shows the results for three types of analysis which involve the performance indexes, the number of hidden neurons used to construct the neural network and also the total simulation time. For the proposed sampling scheme, the performance based on R^2 and d_2 is increased with the increment corresponding to *N*, which approaches 1 for the possible best performance. However, to compromise between the performance and the complexity of the approximate model, for a large dataset, the sample size *N* may be selected at between 25% and 30% from the full dataset. The reason is that no significant improvement on the performance is expected for *N* greater than this value, due to saturation. Notably, the produced *RMSE* and the *MAE* values are relatively small (about ±4% errors) as compared to the maximum output value for the test problem.

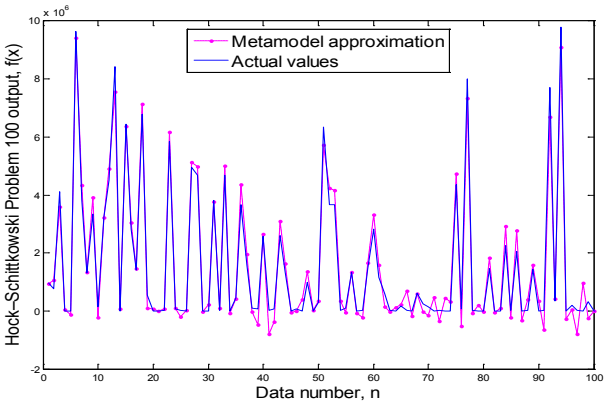

Fig.2. The estimation output for 100 test data using N=1000 (i.e. case no. 3 in Table. 1)

Next, the other two design methods are executed by using the same metamodel configurations. For the *n*-level FFD method, each design variable is assigned with a different number of levels as to generate the different sampling sizes. For example, the [2 3 3 2 3 3 3] configuration will give 972 sampling points' location, which is the product of the number of levels for each dimension. This design approach is in a uniform fashion, by means of a rectangular grid of points. For the LHD technique, a 'maximin' metric, introduced by Johnson *et al.*[20], is considered in this study. This approach yields a randomised sampling plan with projections uniformly-spread onto the axes.

As compared to the *n*-level FFD, at the same sample size, the proposed scheme (i.e. WCD) shows a certain improvement in the size of constructed neural networks, and produces nearly similar performance on the error indications. In the other comparison, the LHD requires about similar network size as WCD, however exhibiting poor performance. Thus, in general, by compromising between the computational cost (i.e. execution time and the network size) and the performance of the model, the WCD method offers a better sampling solution. An example of the estimated output for the case when the sample size is 30 percent of the full dataset is shown in Fig. 2. Therein, the constructed metamodel is able to accu-

Table 1. Metamodel comparison results for Test 1 problem. (Note: full dataset number, N_{full}=4000, sp=4, mse=0.001)

No.	Design name	Details	Sample size, N	% of N	Performance measure				Network size	Simulation time (s)
					RMSE	MAE	R^2	d_2		
1.	WCD		400	10	1.79E06	1.23E06	0.443	0.861	298	42
2.			600	15	1.09E06	7.39E05	0.793	0.948	329	57
3.			1000	25	4.84E05	3.39E05	0.960	0.990	330	93
4.			1400	35	4.18E05	2.98E05	0.971	0.993	333	126
5.			1800	45	3.69E05	2.68E05	0.976	0.994	333	173
6.			2200	55	3.46E05	2.58E05	0.980	0.995	337	250
7.			2800	70	3.14E05	2.32E05	0.983	0.996	341	389
8.			3400	85	3.26E05	2.40E05	0.982	0.995	342	495
9.	n-FFD	[2 3 3 2 3 3 3]	972	24	4.66E05	3.27E05	0.962	0.991	351	99
10.	'n-levels'	[2 3 3 3 3 3 3]	1458	36	3.70E05	2.66E06	0.976	0.994	349	143
11.	LHD	with 'maximin'	1000	25	5.74E05	3.93E05	0.943	0.986	334	89
12.		criterion	1400	35	5.02E05	3.38E05	0.956	0.989	334	127

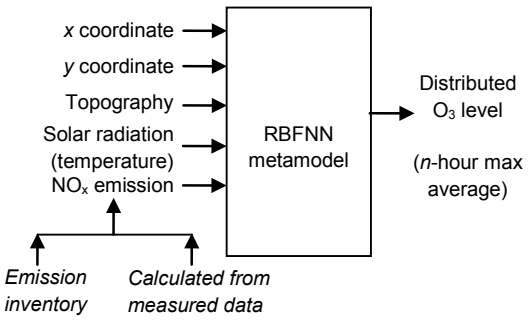

Fig.3. The input-output mapping of the metamodel for spatial estimation of the surface ozone.

rately approximate the true values at most of the points, except for the lower parts (i.e. less than 0 level).

Application: Spatial distribution of air pollutant level

Protecting air quality is essential for the benefits to the public as well as the environment. The air pollution may cause many health problems such as lung disease, asthma and also associated with cancer for some pollutants. It is also affecting the environment by causing some harmful influence to soil, crops, forest, water and wildlife. Thus, authorities are responsible to determine a suitable management policy to protect the air quality. One effective approach is by using data measurement from various monitoring stations across the region. Unfortunately, this is limited to the location of interest. Spatial distribution estimation is an alternative solution to overcome this issue.

In this work, the proposed method of experimental design is applied to spatial estimation for the surface ozone (O_3), recognised as an air pollutant in the tropospheric layer of the atmosphere, of an urban area. This is a very nonlinear and complex estimation task that requires expensive computation. Typically, the dispersion models are used to handle this task, however, they need special software and require a long time in the computational execution. Therefore, a metamodel approach based on neural networks[17] can be used to avoid this complexity as well as to reduce the simulation time.

Spatial estimation model

The functional form of the input-output relationship is not known explicitly because the simulation is a black box. However, from the initial result of a dispersion model, it is suggested that the spatially-distributed ozone levels across the region under consideration is a function of the grid coordinate, topographical information, solar radiation and the ozone's precursor emission, as illustrated in Fig. 3. The background of the problem's design variables will not be discussed in details here (i.e. it is described comprehensively in another parallel work), as this work concerns on the assessment of the sampling designs' performance.

The *x*-*y* coordinates represent the cells' location (in *km*) in *x* and *y* directions. To improve the estimation, topography information is added, consisting of the height information 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 represents a good indicator for proxy variables to the formation of ozone. Basically, there are two important classes of precursors involved in the formation of ozone: volatile organic compounds (VOCs) and Nitrogen Oxides (NO_x), however, only NO_x is to be considered in the modelling as VOCs' concentrations apparently cannot be measured. The NO_x emission rates are added from two sources: the gridded inventory emission rate data, extracted from a photochemical dispersion model (deterministic model), and the calculated emission rates to be converted from the measured data by using the Gaussian dispersion function[21] incorporating the wind speed and the wind direction factors. The network output consists of daily *n*-hour averaged maximum of the ozone concentration (in part per billion, *ppb*), which is extracted from a deterministic model simulation output. Of interest, 1-hour, 4-hour and 8-hour averages are normally selected in the air quality analysis.

Performance analysis

The methodology has been applied to the Sydney basin in New South Wales, Australia. For preparing the dataset, we use the historical data of NO_x that were collected at the monitoring stations around Sydney basin by the Department of Environment in

Table 2. Metamodel comparison results for Test 2 problem. (Note: N_{full}=21000, sp=0.1, mse=0.005)

No.	Design name	Details	Sample size, N	% of N	Performance measure				Network size	Simulation time (s)
					RMSE	MAE	R^2	d_2		
1.	*WCD*	*525*6*	*3150*	*15*	*8.195*	*6.256*	*0.776*	*0.944*	*135*	*70*
2.		*875*6*	*5250*	*25*	*7.953*	*6.121*	*0.789*	*0.947*	*147*	*175*
3.		*1225*6*	*7350*	*35*	*7.814*	*6.011*	*0.797*	*0.949*	*145*	*320*
4.	*n-FFD*	*[3 4 4 4 5]*6*	*5760*	*27*	*12.275*	*8.844*	*0.498*	*0.875*	*115*	*162*
5.	*'n-levels'*	*[4 4 4 4 5]*6*	*7680*	*37*	*10.711*	*7.973*	*0.618*	*0.904*	*121*	*143*
6.	*LHD*	*with 'maximin' criterion*	*5250*	*25*	*15.021*	*10.828*	*0.248*	*0.812*	*114*	*134*
7.			*7350*	*35*	*10.684*	*7.922*	*0.620*	*0.905*	*132*	*287*

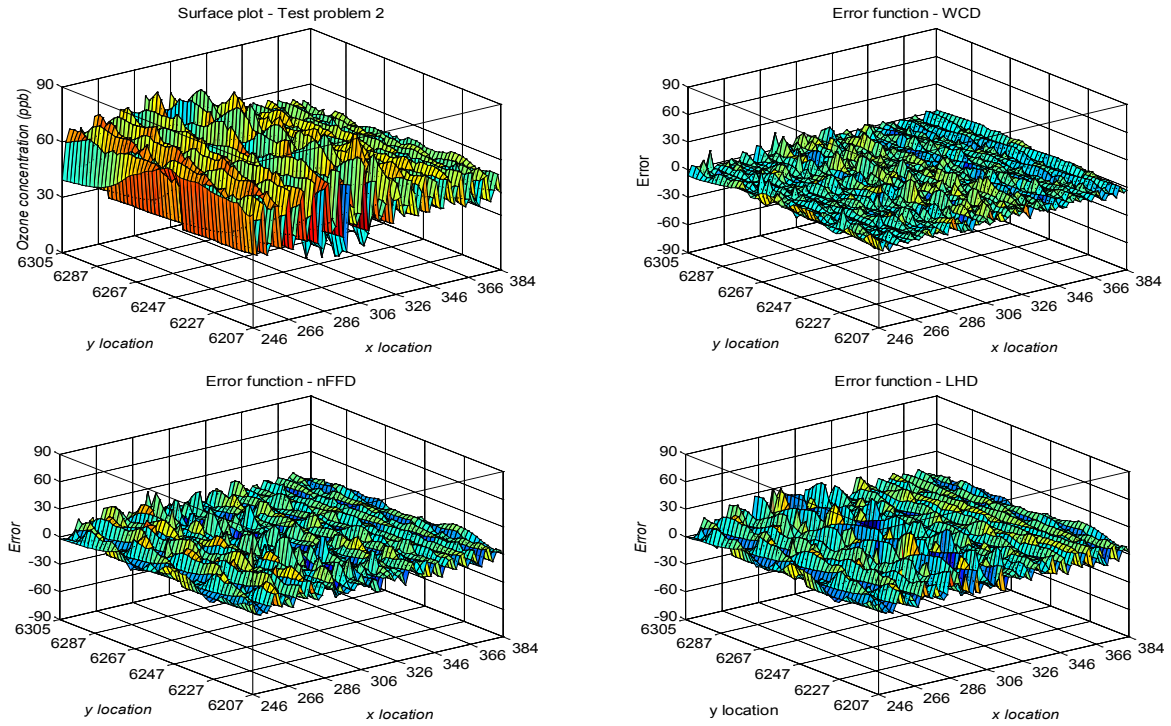

Fig.4 Error analysis of the spatial distribution of the surface ozone concentration

New South Wales. The rests of the input-output data are extracted from the simulation output of The Air Pollution Model and Chemical Transport Model (TAPM-CTM)[22], a pollution model software developed and used in Australia. For demonstration, the 8-hour averaged concentration was used in this work. For each day, the number of datasets corresponding to the number of cells in the studied domain is equal to 3500 (based on a 2*km*×2*km* grid cell). Six episode days are selected to train the metamodel, thus the full input-output dataset consists of 21000 data points.

By using the same steps as for Test 1, the prepared dataset are sampled at different sample size, *N*. Using the RBFNN metamodel with the spread parameter of 0.1 and the *mse* goal of 0.005, the performance of each design method is evaluated. The comparison results are shown in Table 2. To prepare the training dataset, full data points for each day are sampled at *N* size and summed together. As per the performance indexes shown in the table, the proposed sampling method outperforms the other two methods in terms of error criteria. Both n-FFD and LHD method require slightly less computation and smaller network size, however, the produced error indications are very poor (about half of WCD achievement), e.g. in terms of R^2 value. Overall, by compromising the performance and the computational cost, by using the same metamodel design criteria, the proposed approach provides more generalised approximation for air quality modelling.

The surface graphs of the test function and the corresponding error functions using different DOE methods are shown in Fig. 4. The spatially distributed ozone results are obtained for one test day over full cell grids of the domain (i.e. N_{full}=3500 cells). Ideally, the surface graphs of the errors should be flat and near zero. The surface graph of the error function based on the WCD sampling method shows minimal errors when compared with the other two methods. More fluctuated points appear at the left region (i.e. in the west area of Sydney) by using *n*-level FFD and LHD methods.

CONCLUSION

A new method for the sampling design for a neural network metamodel has been presented in this paper. The validity and reliability of the proposed approach has been evaluated in several ways. By using the radial basis function neural network metamodel, the performance of proposed approach was compared with two well-known sampling design strategies; the *n*-level full factorial design and the Latin hypercube design. First, a known non-linear test function, namely, The Hock–Schittkowski Problem 100 was used in the evaluation to validate the effectiveness of the proposed scheme. Next, it has been applied to the air quality problem for the estimation of spatial distribution of the ozone concentration. Using historical meteorological data collected at Sydney's monitoring sites with calibrated input-output dataset from a photochemical dispersion model, the proposed metamodel is capable of predicting the spatially distributed ozone concentration in the interest domain with a fair accuracy. It is also noted that the proposed sampling method outperforms the other two evaluated methods in terms the

network size and the simulation time, for both the test problem and the air quality modelling application.

References

1. Forrester, A.I.J., Sobester, A., and Keane, A.J., *Engineering Design via Surrogate Modeling: A Practical Guide*, Wiley, University of Southampton, UK, 1st edition, 2008.
2. Buragohain, M. and Mahanta, C., "A novel approach for ANFIS modelling based on full factorial design," *Applied Soft Computing*, Vol. 8(1), pp. 609-625, Jan 2008.
3. Swiler, L.P., Slepoy, R., and Giunta, A.A., "Evaluation of Sampling Methods in Constructing Response Surface Approximations," 47th AIAA/ASME/ASCE/AH-S/ASC Structures, *Structural Dynamics and Materials Conference*, No. AIAA-2006-1827, Newport, USA, pp. 1-24, 2006.
4. McKay, M.D., Beckman, R.J., and Conover, W.J., "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code," *Technometrics*, Vol. 42(1), pp. 55-61, Feb 1979.
5. Park, J.S., "Optimal Latin-hypercube Designs for Computer Experiments," *Journal of Statistical Planning and Inference*, Vol. 39, pp. 95-111, 1991.
6. Wang, G.G., "Adaptive Response Surface Method Using Inherited Latin Hypercube Design Points," *Journal of Mechanical Design*, Vol. 125(2), pp. 210-220, June 2003.
7. Abdellatif, A.S., El Rouby, A.B., Abdelhalim, M.B. and Khalil, A.H., "Hybrid Latin Hypercube Designs," In IEEE Xplore: *2010 The 7th International Conference on Informatics and Systems (INFOS)*, pp.1-5, 28-30 March 2010.
8. Fang, H., Rais-Rohani, M., Liu, Z., and Horstemeyer, M.F., "A Comparative Study of Metamodeling Methods for Multiobjective Crashworthiness Optimization," *Computers and Structures*, Vol 83, pp. 2121-2136, 2005.
9. Zhao, D. and Xue, D., "A comparative study of metamodeling methods considering," *Structural and Multidisciplinary Optimization*, Vol. 42, pp. 923-938, 2010.
10. MacKay, D., "Chapter 20. An Example Inference Task: Clustering". *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. pp. 284–292, 2003. Available online.
11. Hamerly, G. and Elkan, C., "Alternatives to the K-means Algorithm that Find Better Clusterings". *Proc. of the 11th Int. Conf. on Information and Knowledge Management (CIKM)*, 2002.
12. Jin, R., Chen, W., and Simpson, T.W., Comparative Studies of Metamodeling Techniques under Multiple Criteria," *Structural and Multidisciplinary Optimization*, Vol. 23(1), pp. 1-13, 2001.
13. Liu, M., Wang, J., and Yan, N., "Design of a Radial Basis Function Neural Network for Attention Tasks Event Related Potentials Extraction," Proc. Int. Conf. on Neural Interface and Control, Wuhan, China, 26-28 May 2005.
14. Ma, L., Xin, K., and Liu, S., "Using radial basis function neural networks to calibrate water quality model," World Academy of Science, Engineering and Technology, Vol. 38, pp. 385-393, 2008.
15. Haykin, S., *Neural Networks a Comprehensive Foundation*, New Jersey, Prentice Hall, 1994.
16. Wahid, H., Ha, Q.P. Duc, H., and Azzi, M., "Estimation of background ozone temporal profiles using neural networks," *Proc. 3rd IEEE Int. Conf. on Intelligence Computing and Intelligent Systems (ICIS 2011)*, Guangzhou, China, pp. 292-297, 2011.
17. Gadner, M.W., and Dorling, S.R., "Statistical surface ozone models: an improved methodology to account for non-linear behaviour," *Atmospheric Environment*, Vol. 34, pp. 21-34, 2000.
18. Meckeshemier, M., Booker, A.J., Barton, R.R., and Simpson, T.W., "Computationally inexpensive metamodel assessment strategies," *AIAA Journal*, Vol. 40(10), pp. 2053-2060, 2002.
19. Hock, W., and Schittowski, K., "Test Examples for Nonlinear Programming Codes," *Springer-Verlag*, Vol. 187, New York, USA, pp. 1-177, 1981.
20. Johnson, M.E., Moore, L.M., and Ylvisaker, D., "Minimax and maximin distance designs," *J. of Satatistical Planning and Infernce*, Vol. 26, pp. 131-148, 1990.
21. Pasquill, F., "Atmospheric dispersion parameters in Gaussian plume modeling," *U.S. Environmental Protection Agency Rep.* EPA-600/4-76-030B, 1976.
22. Hurley, P., "TAPM V4 Part 1: Technical description," *CSIRO Atmospheric Research Paper* No. 25 (2008). Available at: http://www.cmar.csiro.au/research/tapm/docs/tapm_v4_technical_paper_part1.pdf.