Stochastic sampling design for water distribution model calibration

K. Behzadian
Amirkabir University of Technology, Department of Civil and Environmental Engineering, Tehran, Iran
Z. Kapelan
University of Exeter, Centre for Water Systems, Exeter, United Kingdom
D. Savic
University of Exeter, Centre for Water Systems, Exeter, United Kingdom
A. Ardeshir
Amirkabir University of Technology, Department of Civil and Environmental Engineering, Tehran, Iran

ABSTRACT: A novel approach to determine optimal sampling locations under the parameter of uncertainty in a water distribution system (WDS) for the purpose of its hydraulic model calibration is presented. The problem is formulated as a multi-objective optimisation problem under calibration parameter uncertainty. The objectives are to maximise the calibrated model accuracy and to minimise the number of sampling devices as a surrogate of sampling design cost. Model accuracy is defined as the average of normalised traces of model prediction covariance matrices, each of which is constructed from a randomly generated sample of calibration parameter values. To resolve the computational time issue, the optimisation problem is solved using a multi-objective genetic algorithm and adaptive neural networks (MOGA-ANN). The results show that significant computational savings can be achieved by using MOGA-ANN compared to the Monte Carlo Simulation (MCS) model or the GA model based on all full fitness evaluations without significant decrease in the final solution accuracy.

1 INTRODUCTION

The data for calibration of a WDS model is usually collected from a series of field tests at strategic locations within the network, in which pressure heads are recorded (de Schaetzen 2000). The accuracy of calibration is dependant on the quality and quantity of the collected data. Therefore, selection of appropriate locations, called sampling design (SD), has been a challenge among researchers and practitioners especially in recent years (Kapelan et al. 2005a).

Determination of optimal sampling design locations is usually done by evaluating the trade-off between calibrated model accuracy and the cost of sampling design (typically surrogated by the number of sampling devices used). Model accuracy is usually evaluated using some norms of the parameter or the prediction covariance matrix which, in turn, is calculated from the relevant Jacobian matrix (Bush & Uber 1998).

A newly developed model by Kapelan et al. (2003) presented a deterministic multi-objective genetic algorithm (MOGA) for SD with the aim of calibration of WDS models. In the deterministic approach, elements of the Jacobian matrix are calculated prior to the optimisation model run by assuming the model parameter values. This obviously is prone to errors as this kind of information is not readily available. The methodology developed and presented here is trying to overcome this limitation by assuming that each calibration parameter has uncertain value following some pre-defined probability density function.

The assumption of uncertainty in parameters has recently been addressed by a number of researchers in water resources problems (Wu et al. 2006, Kapelan et al. 2005b). Kapelan et al. (2005b) applied the sampling-based technique using Latin hybercube (LH) to deal with uncertainty in parameters. Wu et al. (2006) compared Monte Carlo simple genetic algorithm (MCSGA) with noisy genetic algorithm (NGA) in groundwater sampling network design. They confirmed that NGA can be used as a useful surrogate of MCSGA. However, this approach could still be computationally demanding.

One solution to alleviate this difficulty is to apply meta-models. In a recently developed one, Broad et al. (2005) proposed it as an artificial neural network (ANN) substituting for a complex simulation model of WDS design, in which ANN were trained offline. Yan & Minsker (2006) also developed an adaptive neural network –single objective genetic algorithm model for groundwater remediation design. They saved around 90 percent of the simulation model calls with no loss in accuracy of optimal solutions.

In this paper, a MOGA-ANN algorithm has been developed for the sampling design of a WDS model.

2 OPTIMAL SAMPLING DESIGN

The current SD is carried out under the following assumptions: (1) the type of predicted variables, which include nodal pressure, pipe flows or both, is assumed to be only nodal pressure head; (2) Both nodal demands and pipe roughness coefficients are considered as calibration parameters; (3) the steady-state WDS hydraulic model is calibrated under extended period simulation.

The stochastic SD problem is formulated and solved here as a two-objective optimisation problem under calibration parameter uncertainty. The objectives are to maximise the calibrated model accuracy and to minimise number of sampling devices as a surrogate of sampling design cost.

To quantify the calibrated model prediction accuracy, a first-order second-moment (FOSM) model is used to approximate both parameter covariance matrix and prediction covariance matrix as follows (Bush & Uber 1998, Kapelan et al. 2005a):

$$\mathbf{Cov_a} = s^2.(\mathbf{J^T J})^{-1} \tag{1}$$

$$\mathbf{Cov}_z = \mathbf{J}_z.\mathbf{Cov_a}.\mathbf{J}_z^T \tag{2}$$

Where s=standard deviation of measurement devices; and $\mathbf{J}$=Jacobian matrix of derivatives $\partial y_i / \partial a_k$ $(i = 1,...,N_o; k = 1,...,N_a)$, y=vector of predicted variables in locations of interest, a=vector of calibration parameters, N_o=number of measurement data in both temporal and spatial domains according to measurement locations of interest, N_a=number of calibration parameters; $\mathbf{J}_z$=Jacobian matrix of derivatives $\partial z_i / \partial a_k$ $(i = 1,...,N_z; k = 1,...,N_a)$; z=vector of N_z model predictions of interest, and N_z=number of model predictions of interest in both temporal and spatial domains according to all potential locations of pressure logger installation. The value of the ith diagonal element in matrix $\mathbf{Cov}_z$ indicates the uncertainty of ith model prediction. Therefore, the model prediction uncertainty is presented as the average of all element prediction uncertainties:

$$F_1 = \frac{1}{N_z}\sum_{i=1}^{N_z} \mathbf{Cov}_{z,ii}^{1/2} \tag{3}$$

Since the prediction uncertainty is calculated with the assumption of definite calibration parameter values, the above formula (deterministic approach) can be prone to errors as this kind of information is not definitely? available before model calibration. To remove this limitation, each calibration parameter is assumed here to have uncertain value following some pre-defined probability density function as follows: (1) uncertain pipe roughness coefficient parameters follow a uniform probability density function (PDF) with lower and upper bounds equal to 30% of the deterministic value; (2) uncertain nodal demand parameters follow a Gaussian PDF with coefficient of variation (CV) equal to 0.2.

To deal with the uncertainty of calibration parameter values, noisy fitness function is used here. It has been shown to perform well without sampling a large number of uncertain values (Wu et al. 2006, Gopalakrishnan et al. 2001). Therefore, the first objective value is defined as the average of normalised (relative) traces of model prediction covariance matrices, each of which is constructed from randomly generated sample of calibration parameter values:

$$Max\ f_1 = \frac{1}{N_k}\sum_{j=1}^{N_k} \frac{F_{1,ml}^j}{F_1^j} \tag{4}$$

Where N_k=number of sets of samples; $F_{1,ml}^j$=the value of model uncertainty for ideal state where all potential measurement locations are monitored. This type of calculating the first objective value is called 'full' fitness model henceforth. To do so, N_k sets of uncertain parameter values are randomly generated using LH sampling technique and associated PDFs. The noisy objective value is then calculated by averaging the relative accuracies obtained of running N_k runs of the deterministic SD model. The value of N_k is set to 500 samples that is sufficient for the noisy function based on the performed sensitivity analysis (not shown here).

The second objective value addresses the total cost of sampling. As a surrogate, the number of pressure loggers is introduced as an indicator of sampling cost. Therefore, normalized number of pressure loggers (percentage) is the second objective function. It is presented as follows with its associated constraint:

$$Min\ F_2 = N_p / N_{ml} \tag{5}$$

$$N_p^{\min} \le N_p \le N_p^{\max} \tag{6}$$

Where N_p=number of measurement devices; N_{ml}=number of potential nodes for measurement; $N_p^{\min}$, $N_p^{\max}$= minimum required and maximum number of measurement devices, respectively.

3 METHODOLOGY

The objectives and constraint defined by (4)-(6) indicate a two-objective optimisation problem under uncertainty. However, the calculation of the full fitness model objective (i.e. the model with large number of samples in which the accuracy objective function defined in equation (4) is calculated) involves repetitive calculations of Jacobian matrices, which is usually time-consuming. To resolve the computational time issue, the optimisation problem is solved by using a multi-objective genetic algorithm and adaptive neural networks (MOGA-ANN). Each GA

chromosome is coded as a potential sampling design solution and its fitness is evaluated initially by using the full fitness model. Later on, during the GA search process, the full fitness model is progressively replaced with the periodically (re)trained neural network meta-model where (re)training is done using the data collected by the full model. The ANN is retrained after a pre-specified number of objective function evaluations by the full model. The detailed flowchart of MOGA-ANN is shown in Figure 2.

3.1 *Multi-objective genetic Algorithm*

In this study, a multi-objective evolutionary algorithm known as non-dominated sorting genetic algorithm II (NSGA-II), developed by Deb et al. (2002), is used. NSGA-II alleviates all following difficulties of previous MOGAs: (1) long computational complexity (2) non-elitism approach (3) The need for specification of a sharing parameter. The selection operator in NSGA-II combines the parent and offspring populations in a single population and then selects the best solutions with respect to fitness and spread criteria. NSGA-II can better converge near the true Pareto-optimal front and can better spread solutions through it. More details of this approach can be found in the relevant reference.

Integer value coding is used for the encoding of each chromosome. The number of genes equals the maximum number of measurement devices ($N_p^{\max}$), each of which represents the position of one pressure logger in WDS. A gene with zero value indicates no measurement device is available. When using integer encoding, two or more genes may take the same integer-value values, indicating more than one pressure logger should be installed on the same location. These solutions will be rejected by MOGA due to an increase in cost and no increase in accuracy (Kapelan 2002).

3.2 *Artificial Neural Network (ANN)*

The ANN is used here as a replacement to a full fitness evaluation model used when estimating the model accuracy objective with the idea of making significant computational time savings. However, note that ANN predictions are only approximate and therefore prone to errors in evaluations of objective value. To resolve this drawback, some strategies have been proposed to sample solutions and calculate relevant objective value with full model. Also, the ANNs are periodically retrained within the algorithm progress to improve their prediction accuracy.

Figure 1 shows the architecture of the proposed ANN. As can be seen, a two-layer neural network including input, a hidden and an output layer is assumed. Input data are the potential pressure measurement locations represented by a relevant integer value. Output layer, which has one neuron, is the value of the prediction accuracy objective function defined in equation (4). The second objective function value, i.e. the number of measurement locations, is directly calculated and there is no need to consider it as additional output neuron. In addition, back propagation Levenberg-Marquardt algorithm was used as an ANN training algorithm (Lingireddy & Ormsbee 1998).

3.3 *Main loop*

A flowchart of the proposed MOGA-ANN method is shown in Figure 2. As can be seen, the method is essentially an NSGAII search method which makes use of the artificial neural network and the caching technique. The search process starts by creating the random initial population and evaluating the fitness of each chromosome by using the full model. The data obtained (both chromosome values and the objective function values) is then stored in the cache with the idea of preventing unnecessary, i.e. costly, repetitive fitness evaluations. Note that cache is updated continuously during the search process, i.e. every time chromosome fitness is evaluated using the full model.

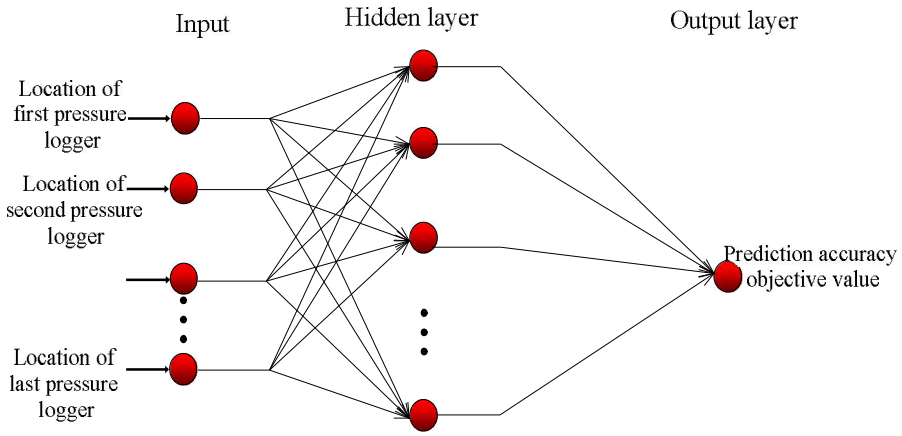

Figure1. ANN Architecture

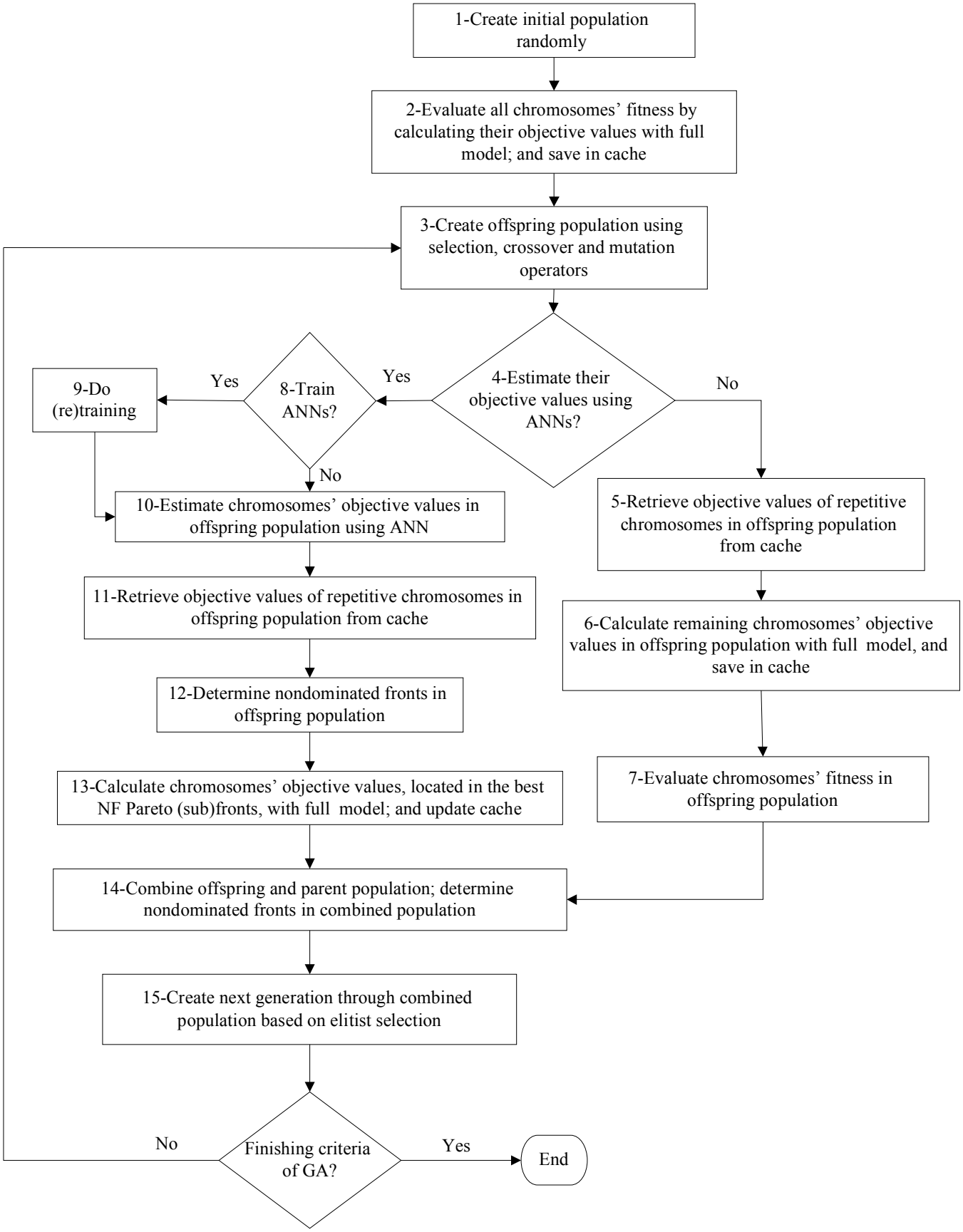

Figure 2. MOGA-ANN Flowchart

The main loop of the algorithm starts with the creation of the offspring population using the NSGA-II selection, crossover, and mutation operators. In the first few generations, chromosome fitness is estimated using the full model only to prepare enough training data for the ANN. Once the ANN is trained for the first time, evaluation of the objective function values is done by using both the ANN and the full model. At first, objective values of all chromosomes in the offspring population are

evaluated using the ANN. Then the offspring chromosomes are compared to the ones previously stored in the cache. If the offspring chromosome is found in the cache then its accuracy objective value (approximated by the ANN) is replaced with the corresponding value from the cache (estimated previously by using the full model).

To improve the algorithm convergence, a (small) number of chromosomes in the offspring population is selected and re-evaluated by using the full model (if it was previously evaluated by the ANN model). The chromosomes selected are the ones present in the best *NF* Pareto (sub)fronts, i.e. subpopulations of the offspring population. Obviously, a trade-off exists here - the larger the *NF* the better from the search accuracy point of view but also the worse from the computational effort point of view. In the case study shown here, the optimal value of *NF* is determined by performing the relevant sensitivity analysis.

Once the offspring population is created by using the above procedure, it is combined with the parent population into a single one. The next generation population is then created by using the standard NSGA-II approach. At this point an additional check is made and if a chromosome is identified with a fitness value estimated by the ANN, its fitness is re-evaluated by using the full model. This is necessary to ensure the good algorithm convergence and it typically involves a small number of chromosomes. The above search process continues until some GA convergence criterion is met (e.g. the pre-specified number of generations).

As an alternative, to calculate objective value of model prediction accuracy in the uncertain environment, an MCS-based model is adopted to compare the results of the optimal sampling locations obtained using noisy objective value to the ones obtained using the MCS method. In the MCS-based model, an equivalent deterministic sampling design optimisation problem (i.e. maximisation of normalised prediction uncertainty defined by (3)) is solved for a number of randomly generated calibration model parameter samples. Based on sensitivity analysis performed, 1000 samples are good enough for MCS model whose statistics sufficiently converge to a unique value. Optimal sampling locations under uncertainty are then determined by identifying the most frequently selected sampling locations in these optimisation runs.

4 CASE STUDY

The above methodology is tested and verified on a literature case study of the Anytown network (Kapelan et al 2003, Ormsbee 1989). The purpose on this case study is to show the capability of the model in decreasing computational effort to get optimal solutions.

Figure 3 shows the layout of Anytown network. The input data has been taken from Ormsbee (1989). Sampling design is performed with respect to calibration parameters of 5 grouped pipe roughness coefficients and 4 grouped nodal demands i.e. the total of N_a=9. All of the network nodes are considered as potential nodes for measurement except for the reservoir and tank nodes, i.e. N_{ml}=16. Full Jacobian matrix J_{ml} is obtained using all potential measurement locations and loading conditions N_O=128 (16 nodes for 8 loading conditions). The standard deviation of all pressure loggers is assumed to be equal to s=0.1m.

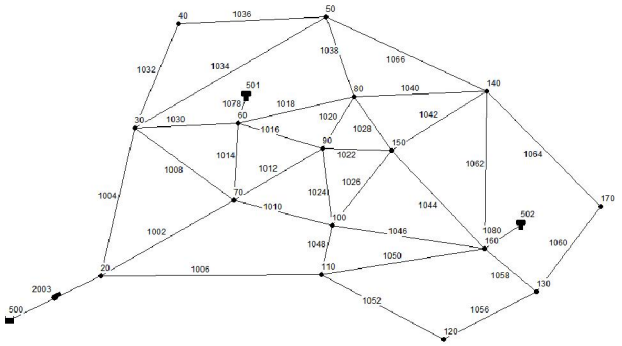

Figure 3. Layout of case study network

5 RESULTS AND DISCUSSION

MOGA model settings were determined after a limited number of trial runs with different initial populations. These parameters used are as follows: population size of 50 chromosomes, binary tournament selection operator, mutation with the probability of 0.25 and one point crossover with the probability of 0.9. All MOGA and MOGA-ANN runs were performed for 500 generations.

The number of best ranked Pareto-(sub)fronts, i.e. subpopulations *NF* was investigated here by performing the sensitivity analysis. The same methodology was used to determine the optimal number of ANN's hidden neurons. The criterion for comparing different settings is the search model reliability denoted here as the percentage of Pareto optimal front points obtained by using the MOGA-ANN model when compared to the full-fitness evaluation based MOGA model. This percentage has been averaged over 20 MOGA runs with different random initial populations to diminish the effect of different search starting points. Figure 4 shows the model reliability with different number of best fronts and different number of hidden neurons. As it can be seen, the reliability of 100% is obtained for *NF* =3 and the optimal number of hidden neurons is 20. The ANN is trained for the first time after 5 generations of full fitness evaluations, and continuously retrained after every

1000 objective function evaluations by the full model (the figure obtained by the sensitivity analysis not shown here).

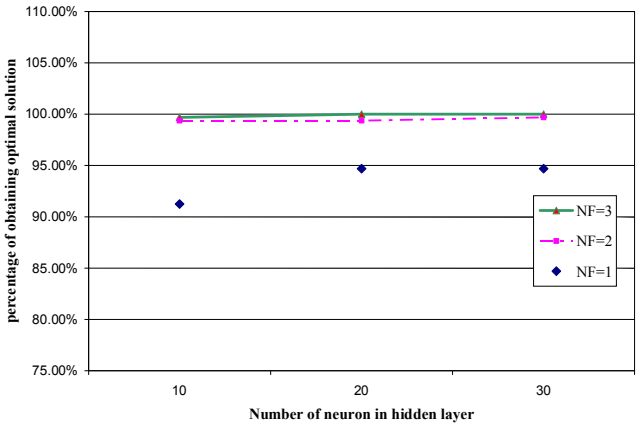

Figure 4. Model reliability for 20 runs with random initial populations; NF=Number of best suboptimal fronts in offspring population, in which the objective value is calculated by full fitness model

After setting the above parameters for the proposed model, the solution of MOGA-ANN as well as MCS-based model were obtained as Pareto optimal fronts shown in Figure 5. Note that for each point on the front, there is a set of optimal locations for installing measurement devices (details presented in Table 1). It can be observed that, when increasing the number of optimal measurement locations to more than 6 or 7 nodes, there is no great improvement in prediction accuracy. Therefore, this point can be introduced as a cost-effective point with regard to one of the selection criteria (Kapelan et al. 2005b).

In Table 1, in addition to the optimal solutions of MOGA-ANN, the percentage of selected sampling locations in the MCS-based model is shown for a given number of monitoring locations. As can be seen, the most frequently selected sampling locations in MCS-based model almost always correspond to the optimal ones in MOGA-ANN. Of course, there are some discrepancies too, in particular in the cases of 3, 4 and 6 monitoring locations. This occurs because of different approaches used in the two methods when dealing with uncertainty. Nevertheless, 97 percent of solutions matched show the similarity in the results obtained using the above two stochastic approaches.

Figure 6 shows the comparison of the number of the actual accuracy function evaluations using the full model, the cache and the ANNs approximations as the MOGA-ANN search progresses. It can be seen that only 12% of chromosomes are evaluated by using the full model. Most of these evaluations occurred in the first five generations of the MOGA-ANN run when the initial ANN training data is collected. After that, the proportion of the full model evaluations is decreasing in the favour of two other means of estimating the solution fitness. The percentage of objective values retrieved from the cache is 25%.

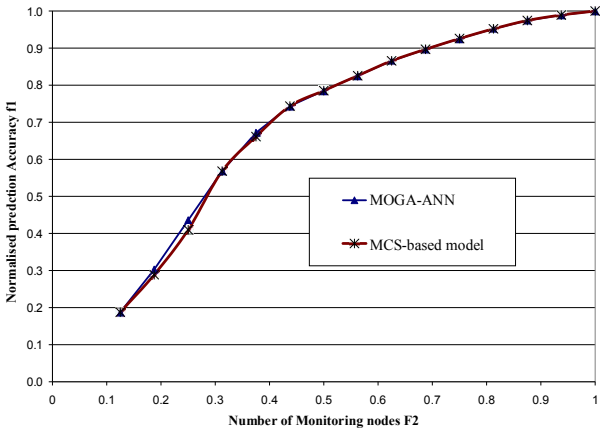

Figure5. Pareto optimal fronts

Table 2 shows the comparison of computational time for different sampling design methodologies (MOGA model with all full model fitness evaluations, the MOGA-ANN model and the MCS model). As it can be seen, the MOGA-ANN method is nearly 9 times faster than the MOGA method based on full model fitness evaluations.

6 CONCLUSIONS

This work proposes an adaptive neural network multiobjective genetic algorithm called MOGA-ANN to determine optimal sampling locations under parameter uncertainty in a WDS for the purpose of its hydraulic model calibration. The ANN is adaptively retrained during the search process. The caching technique was also introduced to efficiently retrieve previously evaluated solutions.

To deal with the uncertainty, noisy fitness function was used in the MOGA-ANN method. Another approach of handling this uncertainty is by using the MCS method. The two methods produced different sets of solutions due to the algorithmic differences. Still, a large proportion of solutions obtained by the two methods were identical.

Table 1. Pareto optimal solutions in MOGA-ANN and percentage of selected sampling locations in MCS-based model

			Network nodes															
F2	f1		20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170
0.13	0.188	Solution	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
		percentage	0	1	11	2	35	3	2	35	5	0	64	22	2	4	3	10
0.19	0.303	Solution	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0
		percentage	0	0	6	1	53	6	1	43	38	17	75	24	1	3	7	25
0.25	0.436	Solution	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1
		percentage	0	0	0	1	58	9	0	69	38	43	89	11	0	1	17	62
0.31	0.568	Solution	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	1
		percentage	0	0	0	1	81	9	1	87	33	73	93	2	0	0	38	81
0.38	0.672	Solution	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	1
		percentage	0	0	1	1	95	17	2	96	43	91	98	2	0	0	62	92
0.44	0.744	Solution	0	0	0	0	1	0	0	1	1	1	1	0	0	0	1	1
		percentage	4	0	5	4	100	34	6	98	54	95	99	9	0	2	92	98
0.50	0.785	Solution	0	0	0	0	1	0	1	1	1	1	1	0	0	0	1	1
		percentage	9	0	10	12	100	43	23	99	59	96	100	44	1	7	98	100
0.56	0.825	Solution	0	0	0	0	1	1	0	1	1	1	1	1	0	0	1	1
		percentage	10	0	18	22	100	60	39	99	69	96	100	74	2	13	99	100
0.63	0.866	Solution	0	0	0	0	1	1	1	1	1	1	1	1	0	0	1	1
		percentage	11	0	26	29	100	76	57	100	82	97	100	92	3	28	100	100
0.69	0.897	Solution	0	0	0	0	1	1	1	1	1	1	1	1	0	1	1	1
		percentage	11	0	42	39	100	84	74	100	91	97	100	99	8	53	100	100
0.75	0.926	Solution	0	0	1	0	1	1	1	1	1	1	1	1	0	1	1	1
		percentage	12	0	63	57	100	90	86	100	97	97	100	100	20	79	100	100
0.81	0.952	Solution	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1
		percentage	14	1	78	81	100	93	94	100	99	97	100	100	46	95	100	100
0.88	0.974	Solution	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		percentage	16	6	97	97	100	99	99	100	100	98	100	100	87	100	100	100
0.94	0.989	Solution	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		percentage	24	76	100	100	100	100	100	100	100	100	100	100	100	100	100	100

"1" means pressure logger should be installed in the node and "0" means no pressure logger is required in the node

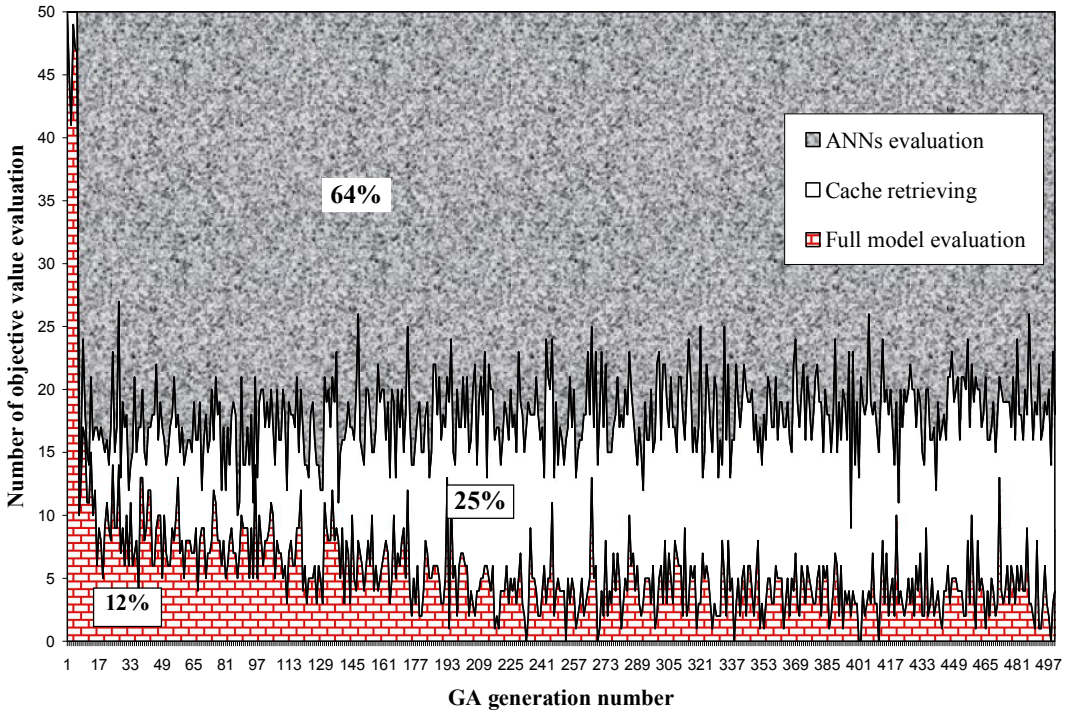

Figure 6. Comparison of objective value evaluations in MOGA-ANN for one sample run

Table 2. Comparison of computational effort to achieve optimal solutions among the models

Model Type	Time (minutes)	The number of deterministic prediction accuracy calculation calls
MOGA	80	12500000
MOGA-ANN	9	1475000
MCS-based	160	25000000

*The number of deterministic prediction accuracy calculation calls for MOGA and MCS-based model is equal to $N_{pop}N_{gen}N_k$, where N_{pop} is GA population size (50 here) and N_{gen} is the number of GA generation before convergence (500 here) and N_k is the number of samples

The results obtained show that large computational savings (90% reduction in CPU time) can be achieved by using the MOGA-ANN when compared to the full-model based MOGA or the MCS model without significant decrease in the final solution accuracy. This finding can be useful in decreasing the computational effort of optimization models with time-consuming fitness evaluations.

7 ACKNOWLEDGEMENTS

The authors wish to acknowledge the support and assistance the first author has received from the British Council (Iran) and Shell for the help provided throughout the research.

REFERENCES

Broad, D. R., Dandy G. C., Maier H. R. (2005), "Water distribution system optimization using metamodels", *Journal of Water Resources Planning and Management*, 131(3), 172-180.

Bush, C. A., and Uber, J. G. (1998). "Sampling Design Methods for Water Distribution Model Calibration." *Journal of Water Resources Planning and Management,* 124(6), 334-344.

Deb, K., Pratap A., Agarwal S., and Meyarivan T. (2002), A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Trans. Evol. Comput*., 6(4), 182–197.

de Schaetzen, W. (2000). ''Optimal calibration and sampling design for hydraulic network models.'' PhD thesis, School of Engineering and computer science, Univ. of Exeter, Exeter, U.K.

Gopalakrishnan G, Minsker BS, Goldberg D. (2001), "Optimal sampling in a noisy genetic algorithm for risk-based remediation design". In: Phelps D, Sehlke G, editors. *Bridging the gap: meeting the world_s water and environmental resources challenges. Proc. world water and environmental resources congress. Washington DC*, ASCE

Kapelan, Z. (2002). ''Calibration of WDS hydraulic models.'' PhD thesis, School of Engineering and computer science, Univ. of Exeter, Exeter, U.K.

Kapelan, Z., Savic D. A., and Walters G. A. (2003) "Multiobjective Sampling Design for Water Distribution Model Calibration", *Journal of Water Resources Planning and Management*, 129(6), 466-479.

Kapelan, Z., Savic D. A., and Walters G. A. (2005a) "Optimal Sampling Design Methodologies for Water Distribution Model Calibration", *Journal of Hydraulic Engineering*, 131(3), 190-200.

Kapelan, Z., Savic D. A., and Walters G. A. (2005b) "Multiobjective design of water distribution systems under uncertainty", *Water Resour. Res.*, 41( ) .

Lingireddy, S., and Ormsbee, L. E. (1998). "Neural networks in optimal calibration of water distribution systems." *Artificial neural networks for civil engineers: Advanced features and applications, I. Flood and N. Kartam,* eds., ASCE, Reston, Va., 53–76.

Ormsbee, L.E. (1989), "Implicit Network Calibration", *Journal of Water Resources Planning and Management*, 115(2), 243-257.

Wu, J., C. Zheng, C.C. Chein, L. Zheng (2006). "A comprative study of Monte Carlo simple genetic algorithm and noisy genetic algorithm for cost-effective sampling network design under uncertainty" *Advance in Water Resources*, 29(1) 899-911.

Yan, S. and Minsker B. (2006). ''Optimal groundwater remediation design using an Adaptive Neural Network Genetic Algorithm.'' *Water Resour. Res.*, 42(5).