UNIVERSITY OF WEST LONDON
The Career University

UWL REPOSITORY

repository.uwl.ac.uk

A comparative study of stochastic and deterministic sampling design for model calibration

Behzadian, Kourosh ORCID: https://orcid.org/0000-0002-1459-8408, Ardeshir, Abdollah, Jalilsani, Fatemeh and Sabour, Farhad (2008) A comparative study of stochastic and deterministic sampling design for model calibration. In: World Environmental and Water Resources Congress 2008, 12-16 May 2008, Honolulu, Hawaii, United States.

doi10.1061/40976(316)482

This is the Accepted Version of the final output.

UWL repository link: https://repository.uwl.ac.uk/id/eprint/10189/

Alternative formats: If you require this document in an alternative format, please contact: open.research@uwl.ac.uk

Copyright: Creative Commons: Attribution 4.0

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

A Comparative Study of Stochastic and Deterministic Sampling Design for Model Calibration

Kourosh Behzadian[1], Abdollah Ardeshir[2], Fatemeh Jalilsani[3], Farhad Sabour[4]

[1] Ph.D. Candidate, Dept. of Civil Eng., Amirkabir Univ. of Tech., Tehran, Iran, Behzadian@aut.ac.ir
[2]Assistant Professor, Dept. of Civil Eng., Amir Kabir Univ. of Tech., Tehran, Iran,
[3]Lecturer, Dept. of Mechanical Eng., Amir Kabir Univ. of Tech., Tehran, Iran,
[4] Director of Sadra Negar Consulting Engineers, Tehran, Iran, Sufarhad@yahoo.com

Abstract

This paper presents and compares two approaches, stochastic and deterministic sampling design, for the purpose of calibrating water distribution system model. Both approaches use a multi-objective genetic algorithm known as NSGA-II to identify the whole Pareto-optimal front of optimal solutions. The relevant objective functions are to maximize the calibrated model accuracy and to minimize the number of sampling devices as a surrogate of sampling design cost. In the deterministic approach, optimal solutions are identified based on the assumed values for calibration parameters. However, the uncertainty of calibration parameters is taken into account in the stochastic approach with some pre-defined probability density functions. Two different stochastic approaches, including noisy fitness function and Monte Carlo simulation, are considered in this study. The efficacy of considering stochastic sampling design rather than deterministic one is assessed by evaluating their objective functions in the simulation of 10000 sampling design problems, each of which is constructed with randomly generated calibration parameters. The stochastic approach is first test on an artificial case study. Then it is applied to a real world water distribution system known as Mahalat model in the central part of Iran. The results of comparison show significant improvements in optimal solutions when using stochastic approaches of sampling design.

Keywords: Stochastic sampling design, Water distribution system, uncertain parameters.

Introduction

Sampling design (SD) for calibration of water distribution system (WDS) models is defined as finding the best locations for installing measurement devices. The collected data from field tests will be used later on for calibration of WDS model (deSchaetzen et al. 2000). The calibration of WDS model is to adjust model parameters so that measured and predicted variables match reasonably. SD problem has been addressed by a number of researchers and practitioners in the last decade (Bush and Uber 1998; Lansey et al. 2001; Kapelan et al. 2003).

When performing SD procedure for the purpose of model calibration, precise values of calibration parameters are unavailable since they will be identified after calibration procedure. Since the access to such precise values is impossible at the time of SD, these values must be estimated based on proposed approximate relationships or manuals. Some researchers such as Bush and Uber (1998) and Kapelan et al. (2003) used assumed

(estimated) values for these parameters. Since SD objective values are associated with the assumption of definite calibration parameter values, they can be prone to errors as this kind of information is not definitely available before model calibration. Lansey et al. (2001) proposed a loop, in which model calibration and SD procedures are repeatedly performed to correct the parameters and SD method. A more comprehensive approach which is used here is to consider uncertainty for these parameters. Therefore, each calibration parameter is assumed here to have uncertain value following some pre-defined probability density function (PDF). In the following sections, two stochastic methodologies of SD are first described. Then it is applied to an artificial case study. Finally the stochastic methodology is applied to a real case study and compared to deterministic approach developed by Kapelan et al. (2003).

Problem formulation

The objective of the SD here is to find a set of optimal pressure measurement locations with the aim of calibrating accurately the WDS hydraulic model. The stochastic SD problem is formulated and solved as a two-objective optimization problem under parameter uncertainty. The two objectives are to maximize the calibrated model accuracy and to minimize the number of sampling devices as a surrogate of sampling design cost. A trade-off between the two objectives is identified for decision making.

Prediction and measurement variables are assumed to be only nodal pressure heads. As a result, if a set of N_l measurement devices with the standard deviation of s are installed in N_l measurement locations of WDS, the variance of predicted variables, denoting prediction uncertainty and obtained by these measurement locations, is estimated as follows (Bush & Uber 1998, Lansey et al. 2001, Kapelan et al. 2005):

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

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

where $\mathbf{J}$=Jacobian matrix of derivatives $3y_i / 3a_k$ $(i = 1,...,N_o; k = 1,...,N_a)$, y=vector of N_o pressure predicted variables in locations of interest, in which pressure loggers are installed, a=vector of calibration parameters, N_o=number of observations, i.e. measurement data in both spatial and temporal domains (e.g. if there are N_t temporal time steps for each of N_l monitoring locations, then N_o=N_t·N_l), N_a=number of calibration parameters, $\mathbf{J}_z$=Jacobian matrix of derivatives $3z_i / 3a_k$ $(i = 1,...,N_z; k = 1,...,N_a)$; z=vector of N_z pressure prediction variables of interest in both spatial and temporal domains. Here spatial domain of N_z is referred to all nodes of WDS. To aggregate the model prediction uncertainty, normalized (relative) prediction accuracy is defined as follows (Kapelan et al. 2003; Bush and Uber 1998):

$$f_1 = \frac{1}{N_z} \prod_{i=1}^{N_z} \mathbf{Cov}_{z,ii}^{1/2} \tag{3}$$

$$F_1 = \frac{f_{1,ml}}{f_1} \tag{4}$$

where $f_{1,ml}$=the value of model uncertainty for ideal state where all potential measurement locations are monitored. To deal with the uncertainty of calibration parameter values, noisy fitness function is used here (Wu et al. 2006). Therefore, the first objective value is defined as the maximization of average of normalized (relative) traces of model prediction covariance matrices, each of which is constructed from randomly generated sample of calibration parameter values:

$$\text{Maximize } F_1 = \frac{1}{N_k} \sum_{j=1}^{N_k} \frac{f^j}{f_{1,ml}^j} \tag{5}$$

where N_k=number of sets of samples and superscript j refers to jth sampling set. This type of calculating the first objective value is called 'full' fitness model henceforth. The second objective value addresses the total cost of sampling. Therefore, normalized number of pressure loggers (percentage) as a surrogate of SD cost is assumed as the second objective value and which defined as follows:

$$\text{Minimize } F_2 = N_l / N_{ml} \tag{6}$$

$$N_l^{\min} :: N_l :: N_l^{\max} \tag{7}$$

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

Multiobjective genetic algorithm (MOGA)

The optimization method which is suited for solving such problems is genetic algorithm (Golberg 1989). As we have two-objective optimization problem, non-dominated sorting genetic algorithm II (NSGAII) developed by Deb et al. (2002) are considered here to be the solver of the optimization problem.

Case #1: Anytown WDS

The first case study is a hypothetical case study known in the literature as "Anytown" WDS model. Figure 1 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/or 4 grouped nodal demands. 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. The standard deviation of all pressure loggers is assumed to be equal to s=0.1m.

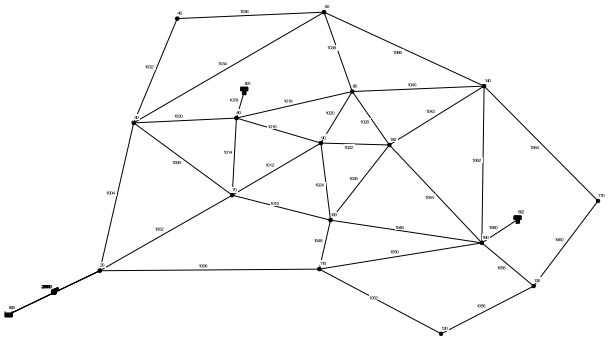

Figure 1. Layout of Anytown case study

Two different approaches for dealing with the uncertainty are as follows: (1) Monte Carlo simulation (MCS) approach; (2) Noisy genetic algorithm (NGA) approach (Wu et al. 2006). The number of samples from calibration parameters and method of sampling should be selected in both approaches. For the first sampling technique, conventional Monte Carlo (MC) sampling technique is considered in which a pre-specified number of equally likely samples of calibration parameters are randomly sampled across a region of the model parameter space. For the second sampling technique, Latin Hypercube (LH) technique is considered in which a pre-specified number of samples are randomly sampled across each region made by dividing equally the model parameter space.

To identify the appropriate number of randomly generated samples in both sampling techniques, a test of sensitivity analysis is performed first for both approaches of MCS and NGA. Thus, the proper number of samples is identified once the statistics of the output converge to unique values (Pasha and Lansey 2005). In MCS approach, the output of stochastic model would be model accuracy objective functions resulted from genetic algorithm optimization. However, in NGA approach, the average of model accuracy objective functions for each possible solution before solving an optimization model can be assumed as the output. The statistics are assumed as average and standard deviation of the output. Therefore, the following steps are done for assessing the convergence of the statistics in stochastic model:

In the MCS approach, (1) randomly generated parameter values of interest are sampled based on a pre-specified probability density function (PDF) for either MC or LH sampling technique. All parameters are assumed to be uncorrelated. (2) The optimization problem (MOGA) of SD model is solved for each realization of model parameters. The model accuracy objective functions for Pareto-optimal front are then calculated (f_1). (3) The average and standard deviation of the model accuracy objective functions are calculated for each corresponding point on Pareto-optimal front for all previous stochastic runs as follows:

$$Ave_i^k = \sum_{j=1}^{i} \frac{f_{1,j}^k}{i} \tag{8}$$

$$STD_i^k = \frac{\sum_{j=1}^{i} (Ave_i^k - f_{1,j}^k)}{i-1} \tag{9}$$

where $f_{1,j}^{k}$ = model accuracy objective function for *k*th point on Pareto-optimal front and *j*th MOGA SD run corresponding to *j*th parameter sampling set; Ave_i^k and STD_i^k =Average and Standard deviation of model accuracy objective function for *k*th point on Pareto-optimal front for all *i* times of MOGA SD runs respectively. (4) Then, the average and standard deviation of all points on Pareto-optimal front are averaged out as follows:

$$X_i = \prod_{k=1}^{m} \frac{Ave^k}{m}_{STD^k} \tag{10}$$

$$S_i = \prod_{k=1}^{m} \frac{i}{m} \tag{11}$$

where X_i and S_i =average and standard deviation of model accuracy objective function for all points on Pareto-optimal front and all *i* times of MOGA SD runs respectively; *m* =number of points on Pareto-optimal front. (5) Steps 2, 3 and 4 are repeated for pre-specified number of runs. It is assumed that the stochastic runs in this study are 1000 times. (6) The above mentioned steps (steps 1 through 5) are performed for 5 random seeds. Finally, the error of average and standard deviation for each number of samples are calculated as follows:

$$error_{ave,i} = \frac{\prod_{j=1}^{n} \left| \frac{\bar{X}_i^j - \bar{X}_{1000}^j}{\bar{X}_{1000}^j} \times 100 \right|}{n} \tag{12}$$

$$error_{STDev,i} = \frac{\prod_{j=1}^{n} \left| \frac{S^j - S_{1000}^j}{S_{1000}^j} \times 100 \right|}{n} \tag{13}$$

where $\bar{X}_i^j$ = average of *i*th samples in *j*th random seeds; $\bar{X}_{1000}^j$ = average of *1000*th samples (the last sample) in *j*th random seed; n =number of random seed; S_i^j = the standard deviation of *i*th samples in *j*th random seeds; S_{1000}^j = the standard deviation of *1000*th samples (the last sample) in *j*th random seed.

In the NGA approach, the process of the assessment is somewhat different. As the output of the NGA approach emerge from the average of model accuracy objective function obtained by stochastic parameters, it does not need to be evaluated after obtaining the optimal solution i.e., solving MOGA optimization problem. Therefore, any typical Pareto-optimal solution can be assessed. The steps required for the assessment are the same as the ones in MCS approach. The only difference is in step 2 in which only model accuracy objective functions for typical Pareto-optimal front are calculated.

Here, based on the above methodology, only the absolute error of the average and standard deviation of the output for NGA approach are show in Figures 2 and 3 respectively. As can be seen, LH sampling technique outperforms MC sampling technique in both average and standard deviation criteria especially in the earlier number of samples. Considering the aforementioned assumption in MCS approach for selecting

the proper number of samples, LH sampling technique is selected. Furthermore, number of 200 samples is enough to achieve a solution with converged statistics.

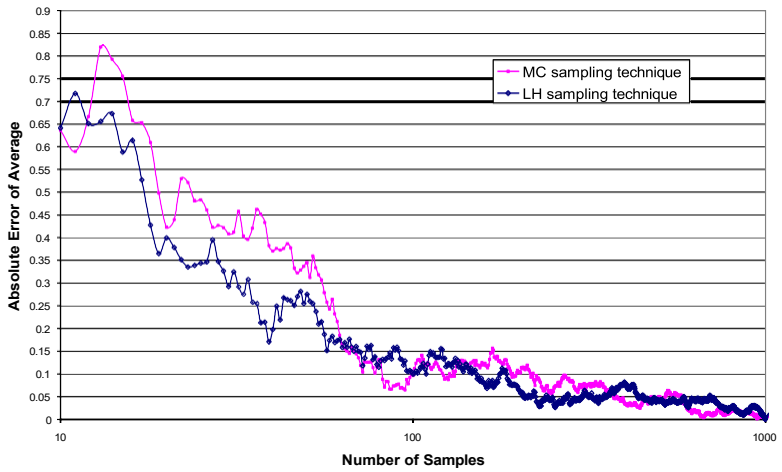

Figure 2. Absolute error of average corresponding to the number of samples for MC and LH sampling technique in the NGA approach

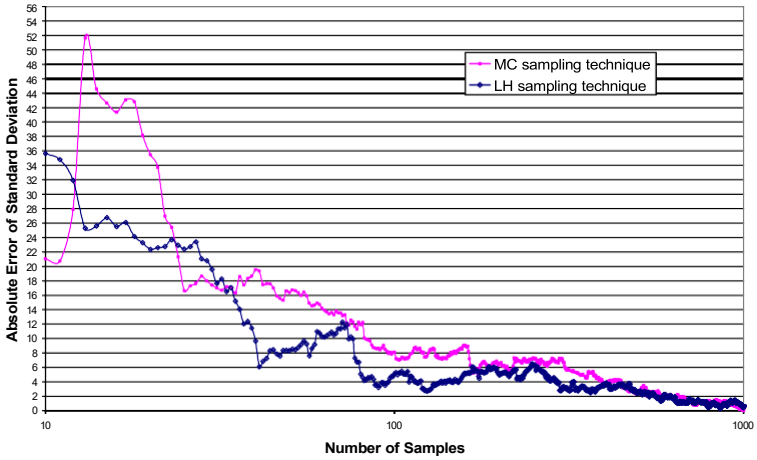

Figure 3. Absolute error of standard deviation corresponding to the number of samples for MC and LH sampling technique in the NGA approach

To perform the two stochastic SD models in this section, the following assumptions are made about uncertain parameters and loading conditions: (1) The calibration parameters are assumed to be both nodal demands and pipe roughness coefficients; (2) uncertain pipe roughness coefficient parameters follow a uniform probability density function (PDF) with lower and upper bounds equal to 30% of the deterministic value; (3) uncertain nodal demand parameters follow a Gaussian PDF with coefficient of variation (CV) equal to 0.2; (4) The hydraulic model is simulated for EPS steady state conditions (8 multipliers).

Considering the above mentioned assumptions for MOGA model and stochastic setting, the Pareto-optimal fronts are obtained by using the two stochastic models. In MCS-based model, optimal locations are determined by identifying the most frequently selected

sampling locations in all deterministic runs with random calibration parameters. Results of Pareto optimal solutions show that the most frequently selected sampling locations in MCS-based model in most cases correspond to the optimal ones in NGA.

To compare the solutions of two stochastic models, the Pareto optimal solutions are then applied to 10000 randomly generated parameter sets. Figure 4 shows the Pareto optimal fronts of two stochastic SD models. The uncertainty of points is shown as the error bars based on the 95% confidence intervals in 10000 parameters realizations. The points also indicate the mean value. As can be seen, the uncertainty of points decreases once the number of monitoring nodes increases. Although the overall trend of both stochastic models including the mean value and 95% confidence intervals is somewhat similar to each to other, NGA outperforms MCS in some points especially for 2 monitoring nodes.

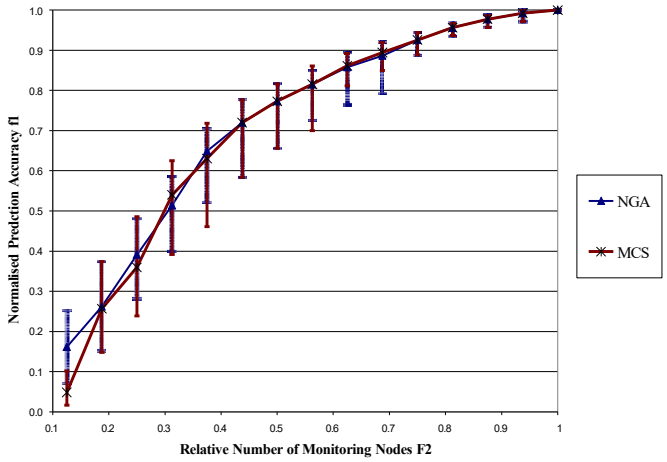

Figure 4. Pareto optimal front for both stochastic SD models including the uncertainty of points on the estimated Pareto optimal front.

The following can be noted: (1) Compared to the deterministic SD, the mean value of uncertainty (standard deviation) in predicting nodal pressure has significantly increased in the stochastic SD. While the uncertainty is less than 0.2 m for all number of monitoring locations in the relevant deterministic SD model, this value is more than 0.2 m for all number of monitoring locations in the stochastic SD model. (2) The value of uncertainty in pressure prediction can be better estimated for each number of monitoring nodes when SD is analyzed under uncertainty. This value would be in a calculated range for optimal measurement locations considering the uncertainty of calibration parameters. For instance, the error (uncertainty) in pressure corresponding to 4 optimal measurement locations is estimated to be between 0.494 m and 0.847 m based on NGA SD analysis. (3) The variation of the uncertainty is dramatically considerable in a few number of monitoring nodes rather than more number of monitoring locations. This may happen because the optimal measurement locations with a few numbers are more sensitive to the changes of calibration parameters. (4) Except for number of monitoring nodes of 2, there is no much difference between the two stochastic SD solutions.

Case #2: Mahalat WDS

The proposed stochastic SD is here applied to Mahalat WDS to verify the capability of the proposed algorithm in a real world WDS case. The city of Mahalat is located in the central part of Iran. The WDS covers approximately 46 km^2, with a population of around 160,000. The city is somewhat steep slope with the lowest point of 1584 meters while the highest one is 1900 meters. Model demands are predominantly domestic with some commercial users. An EPANET hydraulic model (Rossman 2000) was constructed including 1814 pipes, 1771 junctions, 2 tanks, and six PRVs based on the available data. Finally, the skeletonized WDN model was made of 237 pipes and 195 junctions, which is shown in Figure 5. The WDS is supplied by gravity from three wells and two service tanks (reservoirs) around the city. The position of the water supply sources is shown in Figure 5. The average water demand is 158.9 L/S. The water is pumped into the system with a constant rate. The tanks store and balance the fluctuations of water daily consumption (Sadra-Negar 2005).

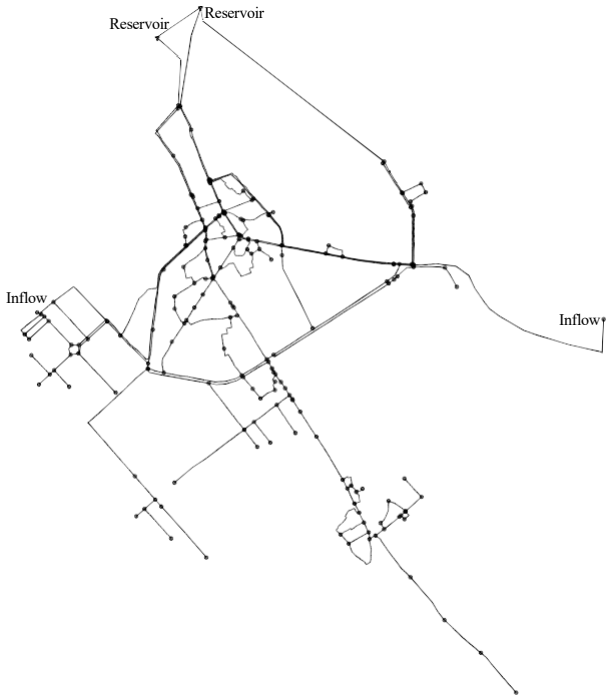

Figure 5. Skeletonized Mahalat WDS model

Further, it is assumed that the WDN model are calibrated for N_a=7 groups. Although there are a large number of pipes (237), number of parameter calibration groups is assumed to be small number because (1) model prediction error will increase if number of calibration parameters increase (2) it was shown that the computational time for running the model will exponentially be enlarged. Grouping was done by dividing the range of HW pipe roughness coefficients into a 7 distinctive ranges. After estimating HW pipe roughness coefficients, their variations were between 78 and 155. Therefore, they have been classified as the ranges of (78, 90], (90,100], …, (130,140] and (140,155]. Then, the average of the HW pipe roughness coefficients in each rang (group) was considered as the representative roughness coefficient of all pipes in that group. In

addition, the model is calibrated for only normal demand loading condition. Note that the standard deviation of all pressure loggers is assumed to be equal to s=1.0 m.

Since the number of calibration parameters is equal to 7 (N_a=7), a minimum number of measurement devices N_{min} constraint equal to 7 is introduced to ensure that the obtained sampling design solution will lead to, at least an over-determined calibration problem. In addition, the whole nodes of the network are considered as potential measurement locations (195 nodes). However, the maximum number of 50 (N_{max}=50) is introduced as SD budget limit, i.e. around 25 percent of the potential nodes.

MOGA model settings were determined after a limited number of trial runs with different initial populations. The following parameters setting have been used for the model: population size of 200 chromosomes, binary tournament selection operator, mutation with the probability of 0.05 and one point crossover with the probability of 0.8. All MOGA runs were performed for 8000 generations. Furthermore, the following assumptions are made about uncertain parameters (1) The calibration parameters are assumed to be only pipe roughness coefficients; (2) uncertain grouped pipe roughness coefficients follow a uniform probability density function (PDF) with lower and upper bounds equal to 30% of the deterministic value; (3) uncertain nodal demand follow a Gaussian PDF with coefficient of variation (CV) equal to 0.3.

The Pareto optimal solutions are obtained by performing two SD models including a stochastic model and a deterministic model. The stochastic models are NGA MOGA SD model (henceforth known as NGA) which is based on noisy fitness function. The deterministic model is standard MOGA SD model (henceforth known as MOGA). To make a better comparison among the solutions of three aforementioned methodologies, they are simulated in an identical uncertain environment. To do so, a total of 10000 SD model simulations were used to estimate the average of relevant accuracy for Pareto optimal solutions. In each SD model simulation, a different set of calibration parameter values was created randomly assuming the uncertainty of parameters as described above.

Figure 6 shows the average of relative accuracy for the three Pareto optimal solutions. This improvement of stochastic SD verifies the superiority of stochastic SD over deterministic SD under uncertainty. It also shows that the solutions obtained by deterministic SD can lead into an average 10% error and maximum 17% error rather than ideal SD, i.e. stochastic SD.

Conclusion

Two stochastic sampling design approaches and one deterministic approach were developed and compared in this paper. All approaches use a multi-objective genetic algorithm known as NSGA-II to identify the whole Pareto-optimal front of optimal solutions. In the deterministic approach, optimal solutions are identified based on the assumed values for calibration parameters. However, the uncertainty of calibration parameters is taken into account in the stochastic approach with some pre-defined probability density functions. First it was applied on an artificial case study. The

stochastic trade-off of the two objective values shows the usefulness of confidence intervals for prediction accuracy of each number of optimal monitoring locations. Then, the comparison of stochastic and deterministic approaches in real WDS shows that the stochastic approach outperforms the deterministic approach especially in large WDS, which can better find optimal measurement locations in uncertain environment of parameters.

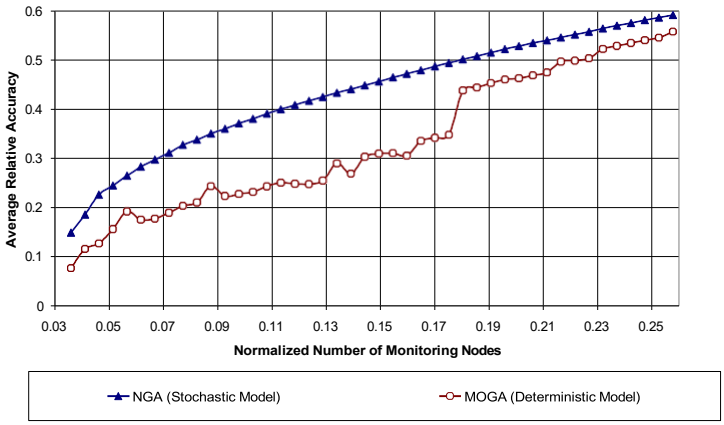

Figure 6. Comparison of stochastic SD and deterministic SD in Malahat WDS model

References

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., Walters, G. A., and Savic, D. A. (2000). "Optimal sampling design for model calibration using shortest path, genetic and entropy algorithms" *Urban Water*, 2, 141–152.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimisation and machine learning*, Addison-Wesley, Reading, Mass.

Kapelan, Z., Savic D. A., and Walters G. A. (2003). "Multi-objective 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. (2005). "Optimal Sampling Design Methodologies for Water Distribution Model Calibration", *Journal of Hydraulic Engineering*, 131(3), 190-200.

Lansey, K. E., El-Shorbagy, W., Ahmed, I., Araujo, J., and Haan, C. T. (2001). "Calibration assessment and data collection for water distribution networks" *J. Hydraul. Eng.*, 127(4), 270–279.

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

Pasha M. F. K., Lansey K. (2005) “Analysis of Parameter Uncertainty on Water Quality in Distribution Systems: Unsteady Conditions”, *Proceeding of CCWI 2005*, UK.

Rossman L. A., *Epanet2 user manual,* US EPA, Washington, D.C., 2000.

Sadra-Negar Consulting Engineers, (2005) “Optimization Study of Mahalat Water Distribution System”, *Water and Wastewater Company of Markazi Province*, Iran.

Wu, J., C. Zheng, C.C. Chein, L. Zheng (2006). “A comparative 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.