Annual Journal of Hydraulic Engineering, JSCE, Vol.50, 2006, February

UNCERTAINTY ASSESSMENT OF HYDROLOGICAL MODEL USING PARETO BASED MULTIOBJECTIVE OPTIMIZATION

Satish BASTOLA [1], Kuniyoshi TAKEUCHI [2] and Hiroshi ISHIDAIRA [3]

[1]Student member, graduate student, Department of Civil & Environmental Engineering University of Yamanashi (4-3-11, Takeda, Kofu, Yamanashi, 400-8511, Japan)
[2]Member of JSCE, Dr of Eng., Professor, Department of Civil & Environmental Engineering, University of Yamanashi (4-3-11, Takeda, Kofu, Yamanashi 400-8511, Japan)
[3]Member of JSCE, Dr of Eng., Associate Professor, Department of Civil & Environmental Engineering University of Yamanashi (4-3-11, Takeda, Kofu, Yamanashi, 400-8511, Japan)

This study proposes the use of multiobjective optimization of a conceptual hydrological model with perturbed data as a sampling method to reproduce the posterior distribution of parameters for the quantification of uncertainty. The Pareto front is found to be sensitive to perturbed data, so model parameters are optimized with different combinations of perturbed data sets to sample behavioral parameters. Latin Hypercube Sampling (LHS) method was used to sample the behavioral parameters in order to evaluate the performance of the proposed method. The performance of simulation for all parameter sets sampled by both methods is evaluated and presented in objective space. The proposed method sampled large parameter sets more efficiently near optimal compared with LHS. The study demonstrates that the Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II) a multiobjective optimizer, with perturbed data set can efficiently explore near optimal parameter space of a conceptual hydrological model.

Key words: *Multiobjective optimization*, *Pareto front, Uncertainty analysis, Hydrological model.*

1. INTRODUCTION

Conceptual hydrological models are often used to predict the relationship between the rainfall and runoff, which is a very complex phenomenon. The parameters of such models are determined by calibration as most of the model parameters are conceptual representations of watershed characteristics. Model calibration allows reducing parameter uncertainty and, therefore, uncertainty in simulation result. Earlier work on automatic calibration of hydrological models suggests that no single objective function is adequate to reproduce different aspects of the hydrograph which led to the formulation of calibration as a multiobjective problem[1), 2), 3)]. Evolutionary algorithms (EAs) have been recognized to be possibly well-suited to multiobjective optimization since they can search for multiple solutions in parallel. All multiobjective EAs clearly outperformed a pure random search strategy [4)].The first pioneering studies on evolutionary multiobjective optimization appeared in the mid-eighties and further advances in EAs were proposed in the years 1993-1994[4), 5)]. Later, these approaches and their variants were successfully applied to various multiobjective optimization problems. The use of Pareto based approaches are growing in hydrological models [3), 6)].

The identification of a best parameter set is necessary for meaningful prediction of flows and parameter regionalization, which however is difficult to realize as measured data for which the calibration is done are not error free and the model never perfectly represents the system. In addition to the identification of best parameter the realistic assessment of parameter uncertainty is essential as multiple parameter sets exist for similar simulations. The first-order approximations to parameter uncertainty near optimum[7)], evaluation of likelihood ratios (Beven and

Binley, 1992) [8] and Markov Chain Monte Carlo Methods[7] are some of the approaches used for the assessment of parameter uncertainty of hydrological model. For nonlinear models, with strong parameter interdependence, first order approximation is quite poor[9].

The objective of this study is to introduce a sampling methodology by coupling multiobjective calibration framework with perturbed data set to efficiently explore optimal parameter space for the reproduction of posterior distribution of parameters to infer parameter uncertainty. The use of large parameter set near optimal to derive the posterior distribution of parameters reduces the parameter uncertainty arising due to suboptimal parameter space.

2. METHODOLOGY

(1) Multiobjective Formulation

The use of single objective to evaluate the simulation can lead to solutions fitting one aspect of the observed hydrograph at the expense of another. To reproduce different aspect of the hydrograph multiple objectives must be considered. To incorporate different objective prior and posterior approaches have been used in the past[1), 2), 3) 6)]. Though prior approaches are simple, they require prior knowledge regarding the problems and lack diversity. Posterior approaches especially Pareto based, though more complex and computationally expensive guarantee convergence to the Pareto optimal front with good population diversity. A general multiobjective optimization problem includes a set of '*m*' decision variables (parameters), and a set of '*q*' objective functions as in Eq. (1). The goal is to

$$Max\ \ F(\beta) = \{ f_1(\beta),....\ f_q(\beta)\} \tag{1}$$

where β (β_1, β_2,. β_m) $\in A$ are the '*m*' parameters, and A is the feasible region . Given two solution β^1, $\beta^2 \in$ A, we say that $\beta^1 = (\beta_1^1,..\beta^1{}_m)$ dominates $\beta^2 = (\beta_1^2,..\beta_m^2)$, If and only if Eq. (2) satisfies, i.e

$$\begin{gathered} f_i(\beta^1) \geq f_i(\beta^2) \forall\ i = 1,.q\ \ and \\ \exists\, i \in \{1,., q\} \text{ with } \ f_i(\beta^1) > f_i(\beta^2). \end{gathered} \tag{2}$$

Pareto based method are posterior approach that uses the concept of Pareto dominance given by Eq. (2) to incorporate different objectives. Genetic Algorithm works with a population of solution which makes them naturally suited to solve multiobjective problems for finding multiple Pareto-optimal solutions. Pareto-based approaches were proposed by Goldberg in 1989 and have become a major focus of Multiobjective genetic algorithm (MOGA) research which was explicitly based on the definition of Pareto optimality. To assure a uniform sampling of the Pareto set, applications of pareto-based MOGA Horn et al. (1994) [4], and Srinivas and Deb (1995) [5] have incorporated niching schemes.

The notion of non-dominating sorting genetic algorithm (NSGA) was first suggested by Goldberg in 1989, and then presented by Srinivas and Deb (1995)[5] for use on multiobjective optimization problems. Based on their findings, Srinivas and Deb (1995) [5] assert that NSGA can tackle higher dimensional and more difficult multiobjective problems. NSGA-II[10] significantly improves upon the original NSGA by (1) invoking a more efficient non-domination sorting algorithm, (2) eliminating the sharing parameter and (3) adding implicitly elitist selection method that greatly aids in capturing high order Pareto surface.

The objective function can be chosen to match an assumption regarding the distribution of the errors present in the observed data. The Root mean square error (RSME) is the maximum likelihood estimator with an assumption that measurement errors are normally distributed and uncorrelated. The Nash Sutcliffe efficiency (NSE) given by Eq. (3), a normalized form of RMSE is commonly adopted for evaluating the simulated hydrograph.

$$NSE = 1 - \frac{\sum_{i=1}^{n} [Q_{obs,i} - Q_{sim,i}]^2}{\sum_{i=1}^{n} [Q_{obs,i} - \overline{Q}_{obs}]^2} \tag{3}$$

where $\overline{Q}_{obs}$ is the average Observed discharge and '*n*' is the number of time step. The different variants of Eq. (3) are considered as objective function to match peak and low flow events[2]. Peak flow NSE is calculated using Eq.(3) for discharge greater than peak flow threshold, and Low flow NSE is calculated using Eq.(3) for discharge less than low flow threshold value. Similarly, the Heteroscedastic maximum likelihood estimator (HMLE) assumes that the measurement errors are normally distributed with zero mean but having Heteroscedastic variance proportional to the observed flows[11]. The normalized form of HMLE is given in Eq. (4).

$$HMLE = 1 - \frac{\frac{1}{n} \sum_{i=1}^{n} w_i (Q_{obs,i} - Q_{sim,i})^2}{\left[\prod_{i=1}^{n} w_i\right]^{\frac{1}{n}} \left(\frac{1}{n} \sum_{i=1}^{n} w_i \left(Q_{obs,i} - \overline{Q}\right)^2\right)} \tag{4}$$

where w_i is the residual at i and is computed as $w_i = \alpha_i^{2(\lambda-1)}$,λ is the variance stabilizing parameter. RMSE tend to emphasize minimization of peak flow error while the HMLE tends to provide more consistent performance across all flow range [1]. Yapo et al. (1998) [1] suggested the use of α_i as observed flow resulted in more stable estimator. In this study four objective functions as mentioned earlier are considered for the calibration of model to consider different aspect of hydrograph.

(2) Description of model

A Tank model with soil moisture structure suggested by Sugawara (1995)[12] for humid basin is used which comprises four vertical tanks with primary storage (PS) and secondary storage (SS) as shown in **Fig. 1**. Water in each tank partially discharges through a side outlet and partially infiltrates through a bottom outlet to the next lower tank. The storage level and the capacity of PS and SS govern the exchange of water between these two storages. A12, A13, A22, A32 and A41 are the runoff coefficient of each tank, A11, A21, and A31 are the infiltration coefficient, H11, H12, H13, H21, H31 are the storage coefficient of each tank, T1 is the coefficients that govern the flow of water between PS and SS. Similarly T2 is the coefficient that governs the flow between PS and Tank-B. A total of 15 parameters are subjected to calibration. SA, SB, SC, SD, Xs, Xp are the initial storage as shown in **Fig.1**. H11 and S1 are the saturation capacity of PS and SS respectively.

To optimize the parameter of Tank model, NSGA-II[10] is used as a optimizer. In NSGA-II, a random parent population of size '*N*' is created which is sorted based on the non-domination. Each solution is assigned fitness equal to its non-domination level. Genetic operators are applied to create an offspring population of same size. Two sets are combined together and sorted according to non-domination. The new population is filled by solutions of different non-dominated fronts starting from the best non-dominated front and continues with solutions of the subsequent non-dominated front. All fronts, which could not be accommodated, are simply deleted, as the size of new population is just half of the size of the combined one. Niching based on the crowding comparison[10] procedure is use to choose the members of the last front which reside in the least crowded region of that front. NSGA-II is capable for higher dimension pareto optimization by generating uniformly distributed pareto front[10].

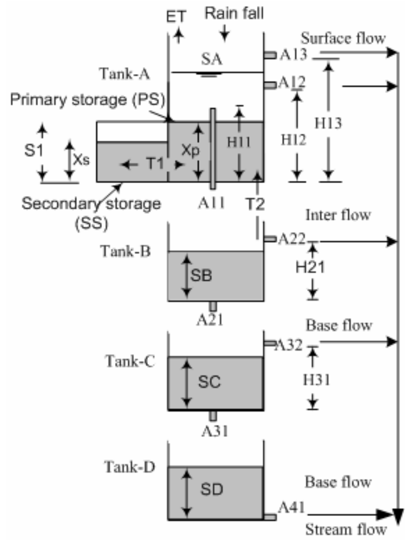

Fig. 1 Schematic diagram for Tank Model

(3) Study area

Uncertainty assessement of the parameters of the Tank model is investigated using historical data (1999-2001) of Ukaibashi, a 487 km^2 sub basin of the Fuji river basin, located in the central part of Japan. The basin lies in an inland region. The mean annual precipitation is approximately 2,100mm. The meteorological data of Mitomi, Kamikane, Kurokoma and Katsunuma stations was used for this study.

(4) Sampling methodology

Uncertainty quantification is essential in numerical models that are used for prediction, which can be obtained by mapping the posterior distribution of parameters. The classical approximation to obtain the posterior probability density function of the parameter uses a first order Taylor series expansion of the nonlinear model equations evaluated at the globally optimal parameter. Using quadratic approximations to the response surface in the region of the best parameters, the multivariate joint probability density function of β (parameter) is then expressed as Eq.(5)[13].

$$p(\boldsymbol{\beta}\,|\,\mathbf{Q}) \ \alpha \ \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_{\mathbf{opt}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_{\mathbf{opt}})\right) \quad (5)$$

Where **X** is a sensitivity matrix evaluated at $\boldsymbol{\beta}_{\mathbf{opt}}$ (optimal parameter). The inferences about **β** can be made assuming the posterior distribution of β is multivariate normal. For rainfall runoff models, such approximation results in poor estimation of posterior distribution of parameters. Beven and Binley[8] have abandoned traditional statistical inference in favor of more general Monte Carlo (MC) based methods. MC based methods are based on the assumption that it is sufficient to have a large number of random sample drawn from a distribution to approximate the form of the density. The LHS and Markov chain Monte Carlo (MCMC) both are efficient sampling methods compared to traditional Monte Carlo methods.

The study proposes an indirect method of sampling the multidimensional parameter space to infer the uncertainty. To conduct a search near global optimum that reveals many behavioral parameter sets resulting similar result, perturbed data sets and multiple objectives are used. The perturbation of input data will lead the search algorithm to explore a large number of points near global optimum, sufficient enough to map the posterior distribution. It is assumed that the input and output uncertainty quantification is given by independent normally distributed random variables. The new perturbed observations are generated by adding a bias drawn from a normal distribution with mean zero and predefined variance to an original value. The general error model is shown in Eq. (6) and (7)

$$\widehat{R}_i = R_i + \varepsilon_i; \quad \varepsilon_i \approx N(0, \sigma_R{}^2) \qquad (6)$$

$$\widehat{Q}_i = Q_i + \zeta_i; \quad \zeta_i \approx N(0, \sigma_Q{}^2) \qquad (7)$$

where $\widehat{Q}_i$ and $\widehat{R}_i$ are the perturbed runoff, and rainfall, Q_i and R_i original runoff and rainfall data respectively. The adopted procedure here after referred as MO, is as follows.

1. Generate the synthetic rainfall and discharge data using Eq.(6) and Eq.(7)
2. The multiobjective optimization with four objective functions NSE, HMLE, Peak flow NSE, and Low flow NSE are used with each of the synthetic data to obtain large sets of Pareto optimal solution.
3. Eq. (8) is used to evaluate the likelihood function as the errors assumed are Gaussian.

$$L(\boldsymbol{\beta}^{\mathbf{p}} \mid \mathbf{Q}) = \exp[-\frac{1}{2}\sum_{i=1}^{n}[\frac{(Q_i - Q_{sim,i})}{\sigma}]^2 \qquad (8)$$

where σ is the standard deviation (STD) of the observed runoff.

4. Eq. (9) is used to compute the profile likelihood ratio. The parameter set satisfying this equation are used to estimate statistical measure of the posterior distribution for each of the parameters.

$$W_p(\theta) = \chi^2_{p,1-\alpha} = -2\log(\frac{L(\boldsymbol{\beta}^{\mathbf{p}} \mid \mathbf{Q})}{L(\widehat{\boldsymbol{\beta}}^{\mathbf{p}} \mid \mathbf{Q})}) \qquad (9)$$

where $\widehat{\boldsymbol{\beta}}^{\mathbf{p}}$ optimal parameter vector. For large samples, under the true parameter value, $W_p(\theta)$ is approximately distributed as a χ^2 random variable with p degree of freedom. The confidence region for **β** consists of all parameter values that would not be rejected at the α significance level.

3. RESULTS AND DISCUSSION

The application of proposed methodology to a hydrologic model is illustrated by calibrating the Tank model using historical data from Ukaibashi basin.

The plot of Pareto front in two dimensions (for convenience) for different perturbed rainfall data is shown in **Fig.2**. Rainfall is perturbed with Gaussian noise of mean zero and STD of 0, 2, 6 and 8 mm/day respectively. It is observed that the perturbation of input data has an impact on the shape of the Pareto front. The Pareto front moved towards higher efficiency side, when data were perturbed with N(0,2) and N(0,6) ,where as the pareto front for N(0,8) fell well below the true (without perturbation) pareto font. Corruption of rainfall with white noise having higher variance offset the relationship between rainfall and runoff which lead the movement of Pareto front on lower efficiency region. It is also observed that due to the inclusion of different objective function the number of Pareto optimal solutions largely increased due to Pareto dominance relationship given in Eq.(2) These two consideration are used to sample parameters in the proposed methodology.

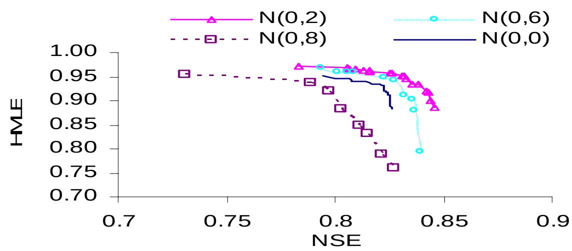

Fig. 2 Effect of perturbed rainfall with Gaussian noise (N (μ, σ)) on Pareto optimal front.

Input error is present when measurements from only a small number of gauges are used, when a more extensive network might be necessary. The variances of rainfall perturbation in Eq. (6) are chosen based on the covariance between rainfall at each gauging station and aerial average precipitation (original value). The rainfall stations are not well distributed in the Ukaibashi basin. The covariance between the aerial average rainfall and spot measurement were observed as high for some year. On the basis of this

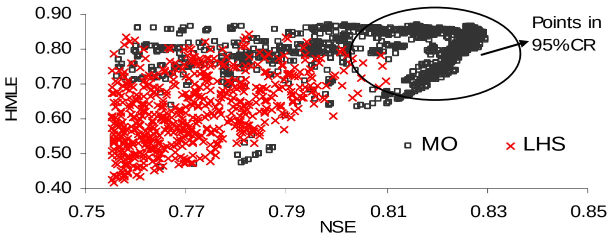

Fig. 3 The objective space for sampled parameter. Rectangles in top-right area are the points sampled by MO where as the cross marks more distributed on lesser efficiency region are sampled by LHS. Points sampled in 95%CR using profile likelihood ratio lie within the circle shown above .The points sampled in 95%CR by MO (~1000) is much higher than those of LHS (~50).

covariance, rainfall data are perturbed with white noise of zero mean and 0, 4, 6, 8 mm/day STD respectively preserving non-negativity in rainfall value using Eq.(6). For runoff 0, 0.5 and 1 m^3/s STD are used to generate synthetic data (Eq. 7) assuming the relative standard error in stream flow measurement is approximately 5%. The total of 12 different combinations of these perturbed rainfall and runoff data are used in this study. In addition to the proposed sampling method LHS, which is a stratified sampling technique is used for comparison. In LHS, the range of probable values for each uncertain input parameter is divided into ordered segments of equal probability. Each parameter is sampled once from each of its possible segments. Once the parameters are sampled by using above mentioned methodology they are again used with original (unperturbed) data to evaluate the model. The performances of the points sampled by both methods are plotted in 2-dimension objective space (NSE & HMLE) for convenience (**Fig.3**). From among all these parameter sets, the parameter set in the 95% confidence regions (CR) is sampled by selecting only those parameters set for which the profile likelihood function (Eq.9) is less than $\chi^2_{p,1-\alpha}$. The number of function evaluations needed to sample this parameter set for LHS doubles that of MO which uses 10^6 function evaluations. It clearly shows the efficiency of MO compared to LHS in sampling parameters near optimal. The plot of ensemble parameter set (normalized) in 95% CR using MO is shown in **Fig. 4**. The variance of NSE and HMLE for all parameter set in 95% CR is less than 0.1%. The parameters in 95% CR with original rainfall are used to obtain the range of simulated hydrograph which is shown in **Fig. 5** and **Fig. 6**. The solid line represents the original observed flow whereas faded area is the range of simulated flow. Around 73% of the wet season flow and 58% of dry

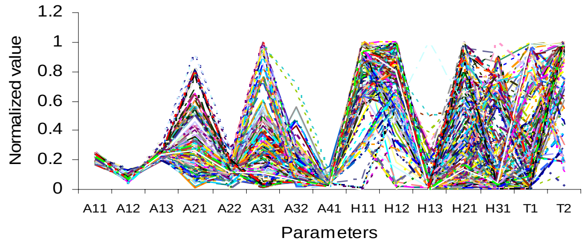

Fig. 4 Sets of normalized parameters lying in 95% confidence region using MO. (Each line represents a single set of normalized parameter)

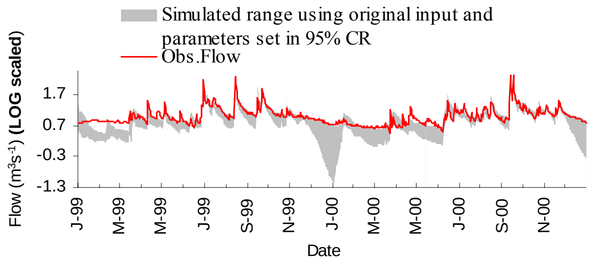

Fig. 5 Hydrograph range for 95% CR for calibration (1999-2000).

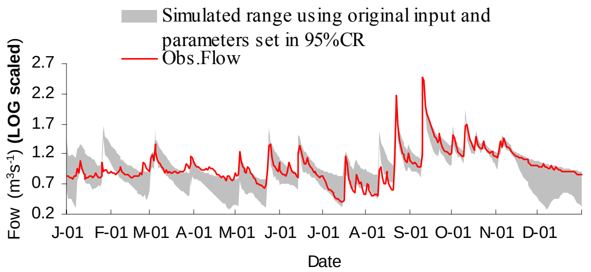

Fig. 6 Hydrograph ranges for 95% CR for validation (2001).

Table1 Statistics of posterior distribution of parameters

Para-meter	MEAN		STD		CV		Opt. Par
	MO	LHS	MO	LHS	MO	LHS	
A11	0.21	0.26	0.02	0.09	0.08	0.37	0.23
A12	0.07	0.09	0.01	0.06	0.20	0.68	0.06
A13	0.24	0.22	0.02	0.08	0.09	0.37	0.27
A21	0.18	0.43	0.13	0.30	0.76	0.68	0.16
A22	0.10	0.17	0.04	0.11	0.40	0.67	0.11
A31	0.40	0.50	0.28	0.30	0.70	0.60	0.32
A32	0.15	0.25	0.12	0.24	0.81	0.97	0.16
A41	0.03	0.14	0.17	0.11	0.52	0.82	0.03
H11	1.83	2.59	0.82	1.16	0.45	0.40	1.08
H12	10.43	13.25	2.48	2.95	0.24	0.22	8.87
H13	34.76	32.22	0.42	1.70	0.01	0.05	34.8
H21	4.18	3.54	0.79	0.87	0.19	0.25	4.93
H31	4.54	3.60	0.67	0.89	0.14	0.25	5.00
T1	0.36	0.50	0.37	0.29	1.02	0.57	0.01
T2	0.79	0.54	0.27	0.28	0.35	0.51	0.98

season flow are within this range for calibration. For validation 78% of wet and 76% of dry season flow are with in this range. **Table 1** shows the statistics of posterior distribution of parameters such as mean, STD, coefficient of variation (CV) of samples generated using MO and LHS, and optimal parameter set (Opt. Par). The Opt. Par is the compromise solution obtained by using global criteria which minimizes the distance to an ideal vector. The runoff and infiltration coefficient of Tank-B and Tank-C (A21, A22, A31, and A32) (**Table 1)** have a higher CV**,** which indicates that these parameters are almost unidentified. High CV for T1 indicates that it is virtually unidentified. The infiltration coefficient A21 of Tank-B is found to be highly correlated (correlation matrix is not shown here) with A22, A31, A32 and H21. In addition, the outlet coefficient of Tank-A and Tank-B are also found strongly correlated. Such a correlation among different parameters leads to numerous identical simulations from a wide range of parameter combination.

Both methods (MO and LHS) reproduced similar first and second moments about origin, where as MO approach was efficient in exploring parameters near optimal. The higher value of CV for LHS compared to MO suggests that the identifiability of parameters can be improved using higher order Pareto optimization and sampling near optimal. The simulation corresponding to posterior mean resulted in 81% NSE for calibration and 91% NSE for validation.

4. CONCULSION

In this paper we have presented a sampling method which uses NSGAII with perturbed input and output data sets for mapping the posterior distribution of parameters of the Tank model near optimal.

Pareto front are sensitive to the perturbation factor used for perturbing the input and output data. The perturbed data can sufficiently explore the parameter space near optimal.

MO is found to be efficient in exploring the near optimal parameter space compared to LHS as MO sampled markedly more parameter sets (~1000) in the 95%CR compared to LHS (~50). In addition, the number of function evaluations required to achieve this large number of parameter sets in the 95%CR was much less for MO compared to LHS. As this method samples more parameter sets near optimal it reduces the uncertainty caused by sub-optimal parameters.

The posterior distribution of parameters for the Tank model at Ukaibashi reveals higher correlation between the parameters, which excludes the possibility of traditional first order approximation to construct posterior distribution. In addition, the higher value of CV for many parameters reveals that the parameters are poorly identified. The proposed method along with being efficient does not use surrogate models like Taylor series approximation of the response surface to infer uncertainty.

ACKNOWLEDGMENT: The authors express special thanks to MEXT and COE program at University of Yamanashi for providing financial support.

REFERENCES

1) Yapo, P.O., Gupta, H. V., and Sorooshian, S.: Multi-Objective global optimization for hydrologic models. *Journal of hydrology,* Vol.204, pp. 83-97, 1998.
2) Madsen, H.: Automatic calibration of conceptual rainfall-runoff model using multiple objectives. *Journal of hydrology*, Vol.235, pp. 276-288, 2000.
3) Fujihara, Y., Tanakamaru, H., Hata, T. and Tada, A.: Performance evaluation of rainfall-runoff models using multi-objective optimization approach. *Proceedings of 2nd Asia Pacific association of hydrology and Water resources Conference,* Vol.2, pp.575-582, 2004.
4) Horn, J., Nafploitis, N., and Goldberg, D.E.: A niched Pareto genetic algorithm for multi-objective optimization. *Proceedings of the first IEEE Conference on evolutionary computation*, pp 82-87, 1994.
5) Srinivas, N and Deb, K.: Multi-Objective function optimization using non-dominated sorting in genetic algorithms. *Evolutionary Computation,* Vol.2, pp.221-248, 1995.
6) Khu, S.T., and Madsen, H.: Multiobjective calibration with pareto preference ordering: an application to rainfall-runoff model calibration. *Water resource research*, Vol.41, W03004 1029 / 2004 Wr003041, 2005.
7) Kuczera, G., and Parent, E.: Monte Carlo assessment of parameter uncertainty in conceptual catchments models: the metropolis algorithm. *Journal of hydrology*, Vol.211, pp. 69-85, 1998.
8) Beven, K.J., and Binley, A.M.: The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, Vol.6, pp. 279-298, 1992.
9) Kuczera, G. and Mroczkowski, M.: Assessment of hydrological parameter uncertainty and the worth of multiresponse data. *Water Resource research*, Vol. 34, pp. 1481–1489, 1998.
10) Dev, K., Agrawal. S, Pratap, A. and Meyarivan, T.: A Fast and elitist multi-objective Genetic Algorithm: NSGA-II, *IEEE transactions on Evolutionary Computation*, Vol.6, pp.182-197, 2002.
11) Sorooshian, S.: Parameter estimation of rainfall-runoff models with heteroscedastic stream flow errors-the noninformative data case. *Journal of Hydrology,* Vol.52, pp.127-138, 1981.
12) Sugawara, M.: Tank model. *In V.P. Singh (ed.) Computer models of watershed hydrology. Water Res. Publ., Littleton, CO*. pp. 165–214, 1995.
13) Box, G.E.P., and G.C, Tiao.: Bayesian Inference in Statistical Analyses, *Addison-Wesley-Longman,Reading,Mass*.,1973.

(Received September 30, 2005)