Journal of Hydroscience and Hydraulic Engineering
Vol. 17, No.1 May, 1999, 11-19

STATISTICAL STUDY OF TANK MODEL IDENTIFICATION BY GENETIC ALGORITHM

By

M. Suzuki, H. Momota, H. Takasaki

Shimizu Corp., Tokyo, 100-0011 Japan

and

K. Jinno, A. Kawamura

Kyusyu Univ., Fukuoka, 812-8185 Japan

SYNOPSIS

The tank model is useful for runoff analysis since it can represent a non-linear stream flow behavior. It is difficult to properly identify a lot of model parameters from observed data. The genetic algorithm (GA) is a search procedure based on the mechanism of natural genetics and is efficient for global optimization. Uncertainty of the identified parameters, which may significantly affect the prediction of stream flow, should be taken into account.

This paper describes the statistics of the identified parameters using GA. The statistics are evaluated by the bootstrap method applied to simulated data fluctuations. Furthermore, the EIC (Extended Information Criterion), which is derived from the bootstrap method, is introduced and applied to select the best tank model.

INTRODUCTION

The in-line four-tank model proposed by Sugawara and others(1) is structurally simple and can represent a non-linear flow behavior. Therefore, it is widely used for long-term runoff analysis. In this model, 16 parameters generally need to be estimated from the volume of runoff in one catchment, and the calibration of the model demands experience. To facilitate calibration, the development of an automation procedure is studied (2),(3).

Studies to estimate parameters by replacing the calibration of the tank model with an optimization problem of a non-linear function, which minimizes the errors in catchment runoff volume, have long been conducted. Kobayashi and Maruyama(4) applied Powell's conjugate direction method to the problem. Watanabe and others(5) suggested to use the Newton's method. Nagai and Kadoya(6),(7) proposed the SP method and the SDFP method incorporating a normalization form into the Powell method and the DFP method. Yasunaga and others(8) tried sequential estimation using the Kalman filter. Since these methods were likely to end in local suboptimum solutions, studies involving global search methods were made. Wang(10), for example, introduced GA into a conceptual rainfall runoff model, and Tanakamaru(11) incorporated GA into a tank model, and both examined applicability. Duan and others(12), and Sorooshian and others(13) proposed the SCE-UA method, which incorporates a GA-like concept into the Simplex method, and applied it to a conceptual rainfall runoff model. Outlines of these studies are given in the paper by Tanakamaru(14).

Parameter estimation in the tank model, if it is regarded as an inverse problem, is an ill-posed problem without uniqueness of the solution or continuity. In this study, therefore, the accuracy and validity of less

subjective analytical methods are examined for the purpose of environmental assessment and in order to obtain an optimum solution for engineering applications which not necessarily coincides with the optimum mathematical solution. Genetic algorithm (GA) suitable for global searches is used as an analytical tool because an objective function that minimizes the errors in catchment runoff volume becomes a multimodal problem of optimization with multiple local solutions. At the same time, the uncertainty of parameters estimated by GA is evaluated by the bootstrap method(15), and the establishment of the optimum number of tanks in the tank model using an information amount criterion is also considered.

ANALYTICAL MODEL AND PARAMETER ESTIMATION

Analytical method

In the development of a runoff analysis model transferring observed rainfall data into runoff volumes, it is crucial to obtain accurate parameters of the model. The tank model is a runoff analysis model frequently used for representing a non-linear flow behavior. The estimation of its parameters is, however, difficult. The target here is, therefore, set at a certain level of estimation achievable with a minimum level of subjectivity. Since GA, which is effective in global searches, handles discrete values, an increase in accuracy requires an increase of bit-strings of parameters. This results in an increase in the number of combinations, and thus an increase in the calculation time required. Here, little reference is attention to GA setting conditions but rather the degree of dispersion of the parameters estimated by GA under certain conditions is examined.

Tank model

In this study, an in-line four-tank model as shown in Fig. 1 is used as runoff analysis model. In the figure, r denotes rainfall (mm/hour), h_i the water depth in the tank (mm), a_i the coefficient of runoff from side hole of the tank (1/hour), b_i the coefficient of seepage from the bottom hole of the tank (1/hour), and c_i represents the height of the runoff hole on the side of the tank (mm).

Then if runoff from the side of the tank, and seepage volume from the bottom of the tank are represented by Q_1 through Q_5 (mm/hour), and by I_1 through I_3 (mm/hour), respectively, the following equations are obtained.

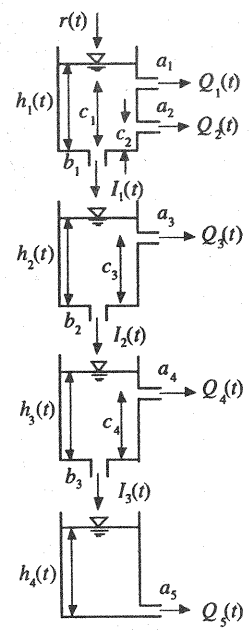

Fig. 1 In-line four-tank model

$$
\begin{aligned}
Q_1(t) &= a_1(h_1(t)-c_1)\ U(h_1(t)-c_1)\\
Q_2(t) &= a_2(h_1(t)-c_2)\ U(h_1(t)-c_2)\\
Q_3(t) &= a_3(h_2(t)-c_3)\ U(h_2(t)-c_3)\\
Q_4(t) &= a_4(h_3(t)-c_4)\ U(h_3(t)-c_4)\\
Q_5(t) &= a_5h_4(t)\\
I_1(t) &= b_1h_1(t)\\
I_2(t) &= b_2h_2(t)\\
I_3(t) &= b_3h_3(t)
\end{aligned}
\qquad (1)
$$

where, U(x) is a unit step function represented by the following equation.

$$U(x) = \begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

Equations of continuity for individual tanks are represented as follows.

$$\begin{aligned} dh_1/dt &= r(t)-Q_1(t)-Q_2(t)-I_1(t) \\ dh_2/dt &= I_1(t)-Q_3(t)-I_2(t) \\ dh_3/dt &= I_2(t)-Q_4(t)-I_3(t) \\ dh_4/dt &= I_3(t)-Q_5(t) \end{aligned} \tag{2}$$

Total runoff Q(t) is represented by the following equation.

$$Q(t) = \sum_{i=1}^{5} Q_i(t) \tag{3}$$

Definition of objective function

For error evaluation, various functions have been used such as those expressed by the following equations. The first shows an error criterion based on least squares, and the second is the logarithm of the first criterion, the third is a chi-square error criterion, and the last one is a relative error criterion. The second and the last equations produce the same criterion.

$$\begin{aligned} J_{LS} &= \frac{1}{M}\sum_{i=1}^{M}\left(Q_c(i) - Q_o(i)\right)^2 \\ J_{LL} &= \frac{1}{M}\sum_{i=1}^{M}\left(\log Q_c(i) - \log Q_o(i)\right)^2 \\ J_{XS} &= \frac{1}{M}\sum_{i=1}^{M}\frac{\left(Q_c(i) - Q_o(i)\right)^2}{Q_o(i)} \\ J_{RE} &= \frac{1}{M}\sum_{i=1}^{M}\frac{\left|Q_c(i) - Q_o(i)\right|}{Q_o(i)} \end{aligned} \tag{4}$$

where, Q_o = observed runoff; Q_c = calculated runoff; M = number of data items.

For functions having a peak such as those for runoff, the last three evaluation criteria can reflect the agreement more accurately than the first criterion based on least squares. Here the second criterion, the logarithm of the criterion based on least squares, is used. No penalty functions such as constraints are considered.

Genetic algorithm (GA)

Described below are methods of searching the parameter vector x (= x_1, x_2, ..., x_n) which minimizes objective function $f(x_1, x_2, \ldots, x_n)$, $x_i^{min} \leq x_i \leq x_i^{max}$, i=1, ..., n, by GA. GA represents x by N strings of symbols using 0 and 1 bits. That is, when handling a continuous quantity, the x_i search range is discretized into 2^L points. The continuous quantity is represented by strings of N=n x L bits, where n times L-bit binary codes are connected. Based on a parameter vector regarding these bit strings as genes, the individual minimizing the objective function is searched by an algorithm con-

sisting of three types of genetic operations, namely, selection, mating, and mutation.

First, the individuals for which the value of the objective function to be minimized is small, are considered to have a large fitness. Selection is a process for random selection of a parent for producing descendants based on a selection probability proportional to the fitness. In this study, what is known as the elite conservation strategy is adopted for selection, in which higher-fitness individuals are conserved for following generations unconditionally. Mating is the creation of a new individual from two selected individuals by exchanging their specific parts. In mutation, a given bit value of an individual selected with a small probability is reversed with a predetermined probability.

APPLICATION OF THE MODEL

Area of application

The data used for the analysis is the daily precipitation averaged over a catchment of the M dam. The analysis was made for five years starting with 1991. Fig. 2 shows daily runoff and precipitation. Snowfall is insufficient to affect runoff. Since this study focuses on flood runoff volume, no evapotranspiration is considered. The four-tank model is applied to simulate the daily runoff volume, and 16 parameters including the initial water depth in the tank are estimated.

Analytical conditions

Upper bounds of search by GA for 16 parameters are defined based on existing references(14) as shown in Table 1. And lower bounds are also defined as 0. The bounds of search are set based on the result of an application of the four-tank model, and are considered large enough for practical purposes.

In GA, the parameter p_i is discretized as shown below.

$$p_i = \Delta p_i z_i$$

$$z_i = (p_i - p_i^{lower}) / (p_i^{upper} - p_i^{lower}) \qquad (5)$$

$$\Delta p_i = (p_i^{upper} - p_i^{lower}) / (2^7 - 1)$$

where, z_i = integer($0 \leq z_1, z_2, \ldots, z_{16} \leq 2^7-1 = 127$); Δp_i = width of discretization of p_i.

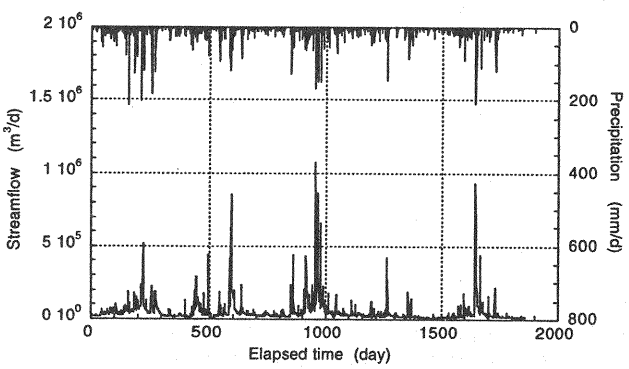

Fig. 2 Daily runoff and precipitation

Table 1 Upper bound of search

Parameter	Upper bound
a_1, a_2, a_3	0.635
a_4	0.127
a_5	0.0127
b_1	0.635
b_2	0.635
b_3	0.127
c_1	127.0
$c_2 \sim c_4$	63.5
$h_{1;0} \sim h_{3;0}$	127.0
$h_{4;0}$	1270.0

Here a GA population of 1,000, a mating rate of 0.6, and a total number of trials of 100,000 are used, and the effects on the search results are studied. The probability of mutation is set at 0.01.

Analytical results

The results of analysis by GA are shown in Table 2. The logarithmic least squares error J_{LL} is represented by σ=0.3105. Parameters obtained by GA, though requiring further improvement, can indicate a general tendency. A comparison between estimation results and observed data is given in Fig. 3.

Differences between observed and estimated logarithmic errors are shown in Fig. 4. They display a normal distribution defined by the objective function, and can be handled as least squares estimates. Here the errors are represented as the units of runoff volume from the tank model. The average value is -0.022, and the variance is 0.101. The autocorrelation function of the logarithmic error is shown in Fig. 5, leaving problems with whiteness. However, the model error is handled as a white noise, and a bootstrapping method is used for the analysis.

Uncertainty of parameters

The error of the estimated parameters in the tank model is evaluated by the objective function defined based on the runoff volume. Here, the error is assumed to be a normal distribution of $N(0.0, \sigma_e^2)$, and the uncertainty of the model parameters estimated by GA is calculated by a bootstrap method. The bootstrap method can create a bootstrap sample by incorporating a model error into time series in estimated runoff volume by a Monte Carlo method, and evaluate its dispersion based on the parameter estimation by GA.

Table 2 Result of analysis

Parameter	Estimated value	Parameter	Estimated value
a_1	0.175	c_1	108.0
a_2	0.120	c_2	15.0
a_3	0.180	c_3	41.0
a_4	0.019	c_4	11.5
a_5	0.0001	$h_{1;0}$	42.0
b_1	0.205	$h_{2;0}$	17.0
b_2	0.490	$h_{3;0}$	97.0
b_3	0.014	$h_{4;0}$	750.0
σ_ε	0.3105		

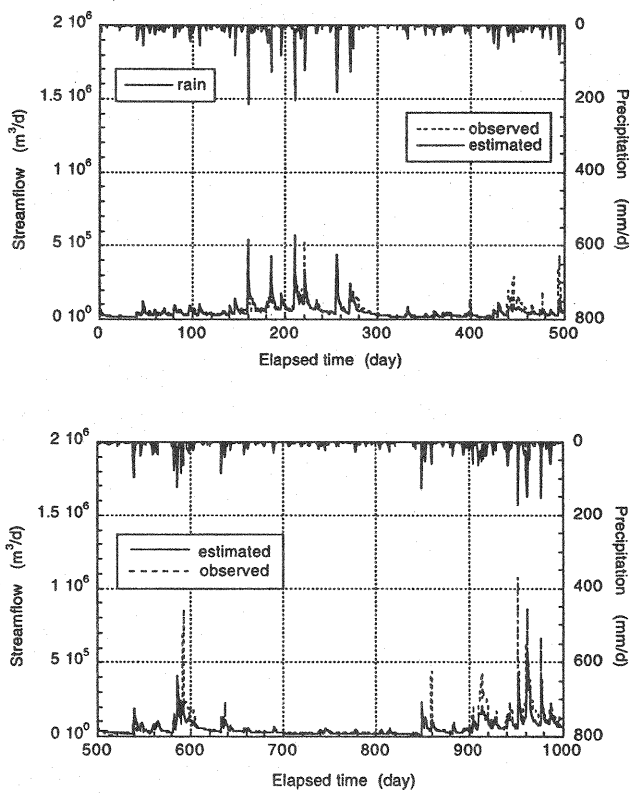

Fig. 3 Estimation results and observed data

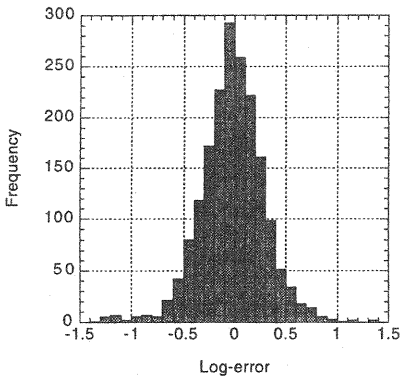

Fig. 4 Histogram of the logarithmic errors

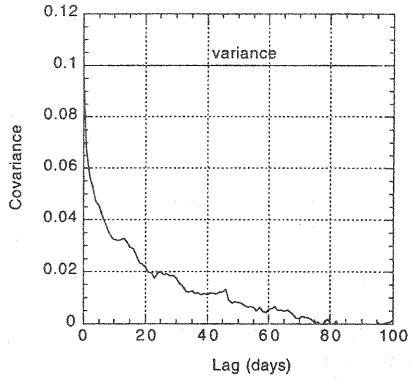

Fig. 5 Auto-correlation of the logarithmic errors

Fig. 6 shows a histogram of a part of the results. The initial water depth in the tank $h_{i;0}$ is uniformly dispersed, and can be an optimum parameter at any values. In conclusion, many parameters are strongly influenced by the search area, and the estimated set of parameters is a non-unique solution.

Parameters obtained by the bootstrap method are expected to be correlated to each other. Such a tendency is expected especially between the water depth in the tank $h_{i;0}$ and the height of the runoff hole on the side of the tank c_i. According to the parameter correlation coefficient shown in Table 3, however, no strong correlation is observed.

Table 3 Parameter correlation coefficients

	$h_{1;0}$	$h_{2;0}$	$h_{3;0}$	$h_{4;0}$	a_1	a_2	a_3	a_4	a_5	b_1	b_2	b_3	c_1	c_2	c_3	c_4
$h_{1;0}$	1.000	-0.226	-0.271	0.024	-0.045	0.017	0.063	-0.136	-0.078	-0.211	0.034	-0.194	0.017	0.128	0.204	0.151
$h_{2;0}$	-0.226	1.000	-0.375	-0.045	-0.041	-0.090	0.020	-0.126	0.109	-0.031	-0.080	-0.119	-0.049	0.198	-0.018	0.252
$h_{3;0}$	-0.271	-0.375	1.000	-0.094	0.034	-0.006	0.064	-0.152	0.144	-0.112	0.164	-0.079	0.149	-0.055	0.038	0.377
$h_{4;0}$	0.024	-0.045	-0.094	1.000	0.000	-0.213	0.044	-0.117	-0.256	-0.141	-0.008	0.103	-0.024	-0.028	0.139	0.043
a_1	-0.045	-0.041	0.034	0.000	1.000	0.232	0.111	0.162	0.250	0.150	-0.110	-0.051	0.409	-0.192	-0.071	0.000
a_2	0.017	-0.090	-0.006	-0.213	0.232	1.000	0.084	0.546	0.498	0.440	-0.236	0.068	0.198	-0.226	-0.171	-0.075
a_3	0.063	0.021	0.064	0.044	0.111	0.084	1.000	0.224	0.195	-0.201	0.225	0.043	0.096	0.097	0.460	0.186
a_4	-0.136	-0.126	-0.152	-0.117	0.162	0.546	0.224	1.000	0.687	0.318	-0.241	0.419	-0.165	0.017	0.066	-0.015
a_5	-0.078	0.109	0.144	-0.256	0.250	0.498	0.195	0.687	1.000	0.274	-0.166	0.239	-0.097	0.119	0.028	0.463
b_1	-0.211	-0.031	-0.112	-0.141	0.150	0.440	-0.201	0.318	0.274	1.000	-0.337	0.074	-0.113	-0.259	-0.678	-0.219
b_2	0.034	-0.080	0.164	-0.008	-0.110	-0.236	0.225	-0.241	-0.166	-0.337	1.000	-0.138	0.083	0.094	0.139	0.090
b_3	-0.194	-0.119	-0.079	0.103	-0.051	0.068	0.043	0.419	0.239	0.074	-0.138	1.000	0.074	-0.145	0.228	-0.307
c_1	0.017	-0.049	0.149	-0.024	0.409	0.198	0.096	-0.165	-0.097	-0.113	0.083	0.074	1.000	-0.654	0.079	-0.053
c_2	0.128	0.198	-0.055	-0.028	-0.192	-0.226	0.097	0.017	0.119	-0.259	0.094	-0.145	-0.654	1.000	0.230	0.255
c_3	0.204	-0.018	0.038	0.139	-0.071	-0.171	0.460	0.066	0.028	-0.678	0.139	0.228	0.079	0.230	1.000	0.178
c_4	0.151	0.252	0.377	0.043	0.000	-0.075	0.186	-0.015	0.463	-0.219	0.090	-0.307	-0.053	0.255	0.178	1.000

CRITERION FOR MODEL SELECTION

AIC (Akaike's Information Criterion), one of the most well-known model selection criteria, uses an average logarithmic likelihood, which is obtained by a maximum logarithmic likelihood according to the number of parameters. The model which minimizes AIC is considered appropriate. AIC is, however, a criterion based on maximum likelihood estimates, and thus applicable only by a maximum likelihood method. EIC (Extended Information Criterion), an extension of AIC, correctes biases in logarithmic likelihood according to the data by a bootstrap method. It is time consuming but has a large applicability(16). It can be expressed by the following equation.

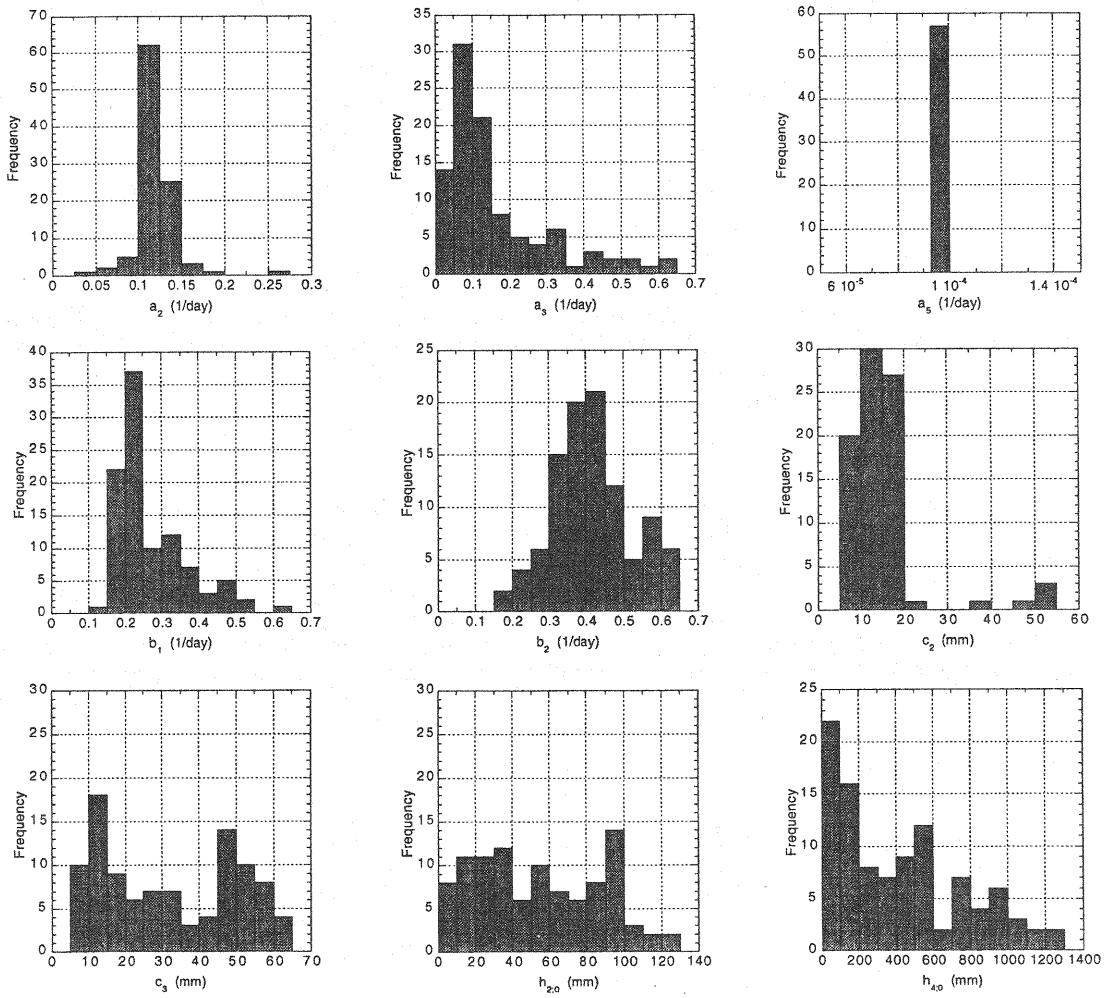

Fig. 6 Histograms of a part of the results

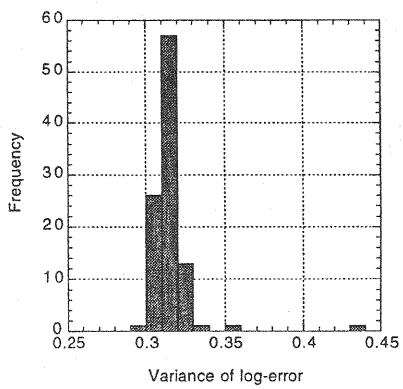

Fig. 7 Histogram of the objective function errors of the four-tank model

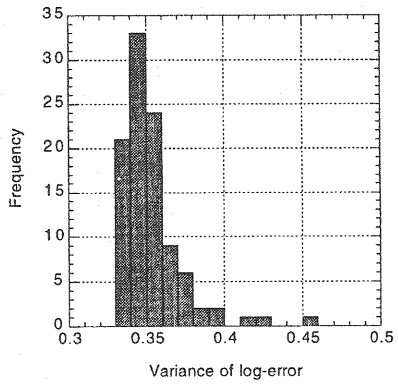

Fig. 8 Histogram of the objective function errors of the three-tank model

$$EIC = -2 \log f(X \mid \hat{\theta}) + 2E_x\left[\log f(X^* \mid \hat{\theta}^*) - \log f(X \mid \hat{\theta}^*)\right] \qquad (6)$$

where, X^* = a bootstrap sample; θ = an estimated parameter vector.

Using EIC, the four-tank model shown in Fig. 1 and a 12-parameter in-line three-tank model, created by removing the third tank from the top from the four-tank model, are compared. Histograms of the objective function errors of the four- and three-tank models are shown in Figs. 7 and 8, respectively. Smaller dispersion is observed for the four-tank model. As a result of comparison in the estimated error between estimated three- and four-tank models, it was found that the first term on the right side of the above equation was larger, and the correct term, the second on the right was smaller for the three-tank model than for the four-tank model. Finally, the parameter vector was 498.87 for the four-tank model, and 1002.15 for the three-tank model when EIC was used. Thus the four-tank model was considered superior appropriate and selected.

CONCLUSIONS

Calibration of a runoff analysis model by GA was conducted in a catchment of the M dam, and the statistical uncertainty of the model was evaluated. An in-line four-tank model was used as a catchment runoff analysis model, and estimation of its parameters without prior information was performed. As a result, it was possible to obtain an optimum solution for engineering use, and identify its uncertainty although it is a multimodal ill-posed problem without uniqueness or solution continuity. Problems left to be tackled include the dispersion of identified parameters and the definition of the objective function.

REFERENCES

1. Sugawara, M. : Method of runoff analysis, Kyoritsu Shuppan Press, 1972. (in Japanese)
2. Sugawara, M., E. Ozaki, I. Watanabe and Y. Katsuyama : Method of automatic calibration of tank model (first report), Research Notes of the National Res. Center for Disas. Prev., Japan, No.17, pp.43-89, 1977. (in Japanese)
3. Sugawara, M., I. Watanabe, E. Ozaki, Y. Katsuyama : Method of automatic calibration of tank model (second report), Research Notes of the National Res. Center for Disas. Prev., Japan, No.20, pp.157-216, 1978. (in Japanese)
4. Kobayashi, S. and T. Maruyama: Search for the coefficients of the reservoir model with the Powell's conjugate direction method, Trans. JSIDRE, No.65, pp.42-47, 1976. (in Japanese)
5. Watanabe, K., K. Tateya, K. Matsuki and K. Hoshi: Refinements to parameter optimization in the tank model, Proceedings of the 33rd Japanese Conference on Hydraulics, pp.55-60, 1989. (in Japanese)
6. Nagai, A. and M. Kadoya : Optimization techniques for parameter identification of runoff model, Annuals, Disas. Prev. Res. Inst., Kyoto Univ., No.22B-2, pp.209-224, 1979. (in Japanese)
7. Nagai, A. and M. Kadoya : Storage model for analyzing flood and long term runoff and its optimum identification, Annuals, Disas. Prev. Res. Inst., Kyoto Univ., No.26B-2, pp.261-272, 1983. (in Japanese)
8. Yasunaga, T., K. Jinno and A. Kawamura: Change in the runoff process into an irrigation pond due to land alteration, Proceedings of the 36th Japanese Conference on Hydraulics, pp.629-634, 1992. (in Japanese)
9. Kadoya, M., H. Tanakamaru, A. Nagai and M. Kaneguchi: Application of the long- and short-term runoff model in the upper area of the river Echi and real-time flood forecasting, Annual Reports of Water Re-

sources Research Center, DPRI, Kyoto Univ., No.9, pp.45-60, 1989. (in Japanese)
10. Wang, Q.J.: The genetic algorithm and its application to calibrating Conceptual Rainfall-Runoff Models, Water Resour. Res., Vol.27, No.9, pp.2467-2471, 1991.
11. Tanakamaru, H.: Parameter identification of tank model with the genetic algorithm, Annuals, Disas. Prev. Res. Inst., Kyoto Univ., No.36B-2, pp.231-239, 1993. (in Japanese)
12. Duan, Q., S. Sorooshian and V.K. Gupta: Effective and efficient global Optimization for Conceptual Rainfall-Runoff Models, Water Resour. Res., Vol.28, No.4, pp.1015-1031, 1992.
13. Sorooshian, S., Q. Duan and V.K. Gupta: Calibration of rainfall-runoff models : Application of global optimization to the Sacramento soil moisture accounting model, Water Resour. Res., Vol.29, No.4, pp.1185-1194, 1993.
14. Tanakamaru, H. and S.J. Burges: Application of global optimization to parameter estimation of the tank model, Proc. of the Int. Conf. on Water Resour. & Environ. Res., Vol.II, pp.39-46, 1996.
15. Efron, B.: Bootstrap methods: Another look at the jackknife, Annals of Statistics, Vol.7, No.1, pp.1-26, 1979.
16. Ishiguro, M., Y. Sakamoto and G. Kitagawa: Bootstrapping log likelihood and EIC, an extension of AIC, Annuals of the Inst. of Statistical Mathematics, Vol.49, p.411-434, 1997.

(Received October 5,, 1998 ; revised March 8, 1999)