ARTIFICIAL NEURAL NETWORK FOR PREDICTION OF WATER QUALITY IN PIPELINE SYSTEMS

Ju-Hwan Kim and Jae-Heung Yoon

Water Resources Research Institute, Korea Water Resources Corporation, Jeonmin-Dong 462-1, Yusung-Gu, Daejeon, Korea

Abstract: The applicabilities and validities of two methodologies for the prediction of THM (trihalomethane) formation in a water pipeline system were proposed and discussed. One is the multiple regression technique and the other is an artificial neural network technique. There are many factors which influence water quality, especially THMs formations in water pipeline systems. In this study, the prediction models of THM formation in water pipeline systems are developed based on the independent variables proposed by American Water Works Association(AWWA). Multiple linear/nonlinear regression models are estimated and three layer feed-forward artificial neural networks have been used to predict the THM formation in a water pipeline system. Input parameters of the models consist of organic compounds measured in water pipeline systems such as TOC, DOC and UV254. Also, the reaction time to each measuring site along pipeline is used as input parameter calculated by a hydraulic analysis. Using these variables as model parameters, four models are developed. And the predicted results from the four developed models are compared statistically to the measured THMs data set. It is shown that the artificial neural network approaches are much superior to the conventional regression approaches and that the developed models by neural network can be used more efficiently and reproduce more accurately the THMs formation in water pipeline systems, than the conventional regression methods proposed by AWWA.

Keywords: pipeline system, disinfection by-product, THM formation, neural network, regression method

1. INTRODUCTION

New restrictive rules for filtration of surface water and maximum levels of total trihalomethane in pipeline systems are being imposed by the drinking water standards. The new disinfection/disinfectants by-products (DBP) rule addresses the possibility of specifying maximum contaminant levels for each individual THM species since their health risks differ significantly. The study by US national cancer institute in 1976 indicated that chloroform, a major component of THMs, is an animal carcinogen and eventually a suspected human carcinogen. Bromoform and bromo-dichloromethane were reported carcinogenic later. It is clear that prop-

erly developed water quality models to simulate the temporal and spatial variations of different substances in pipeline system can potentially assist the water utilities' operators in abiding with the drinking water quality standards. A number of such kind of models have emerged during last decade. They were mainly developed to model the chlorine under different dynamic conditions. However, the appearance and monitoring of trihalomethane in water pipeline systems is practically difficult due to the complexities of analysis.

The THM compounds develop in chlorinated water containing organic precursors, such as humic and fulvic acids. It has been reported that the relative contribution to the formation of THMs by the humic acids react more readily with the chlorine. Even though it is well known that TTHM increases with time, information about the reaction of the mechanism of the formation of THM and its species is still limited. This paper presented modeling techniques for the prediction of , multiple linear and nonlinear regression and artificial neural networks (ANN) and application results for the prediction of THM formation in water distribution system based on statistical analysis of the observed water quality data in water distribution system. The model parameters are selected as used in formula proposed by American Water Works Association (1993), Urano (1983) and Engerholm-Amy (1987). A steady state hydraulic analysis program (KYPIPE model) was applied to get hydraulic characteristics of the water distribution system under investigation.

2. BACKGROUND INFORMATION

In this section, two approaches of relevant prediction methods are reviewed. The prediction of THM formation in distribution systems is influenced by other water quality parameters. Therefore, prediction methods that have been developed for THM formation are also useful for regression method based on statistical analysis of observed data.

There is a linear relationship between chlorine consumption and the production of THMs with a reaction, and the THM evolution is shown to be the function of many water quality parameters. In this study, the relationships between THM formation and each water quality parameter in water distribution system are investigated and analyzed. And the prediction models of THM formation are developed as a function of water quality parameters, including the total organic carbon, type of organic precursors, pH of chlorination, temperature, UV light absorbance, bromide level, and reaction time. The formulas are expressed as a function of the above water quality parameters using multiple linear and nonlinear regression procedures.

Also, artificial neural networks are introduced and its theoretical background will be provided and discussed for the present study. The calculated results from a KYPIPE model for hydraulic pipe flow analysis were used to get reaction time.

2.1 Multiple linear and nonlinear regression model

This approach makes a description by using a set of equation. Information required to construct this model can be obtained in variety ways, such as observations and experiments. Instead of kinetics of THM formation, other water quality parameters in distribution systems have relative contribution to the formation of THM. Regression equation can be generated by parameters information. Scatter diagrams of each of the

predictor variables were individually analyzed against the response variable. Inclusion of more predictor variables in a multiple linear regression model is worth testing in this study. A multiple linear regression model that expresses the relationship between THM formation and water quality parameters as independent variables, is given by equation (1). And nonlinear regression model can be described by equation (2).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

$$y = \beta_0 \cdot x_1^{\beta_1} \cdot x_2^{\beta_2} \cdots x_n^{\beta_n} \quad (2)$$

where $\beta_0, \beta_1 \cdots, \beta_n$ denote model parameters, $x_1, x_2, \cdots, x_n$ and y denotes independent and dependent variables, respectively.

2.2 Artificial neural network

An artificial neural network(ANN) is a network of parallel, distributed information processing systems that relate an input vector to an output vector. It consists of a number of information processing elements called neurons or nodes, which are grouped in layers. The input layer processing elements receive the input vector and transmit the values to the next layer of processing elements across connections where this process is continued. This type of network, where data flow on way(forward), is known as a feed-forward network. A feed-forward ANN has an input layer, an output layer, and one or more hidden layers between the input and output layers. Each of the neurons in a layer is connected to all the neurons of the next layer, and the neurons in one layer are connected only to the neurons of the immediate next layer. The strength of the signal passing from one neuron to the other depends on the weight of the interconnections. The hidden layer enhances the network's ability to model complex functions. The data passing through the connections from one neuron to another is manipulated by weights that control the strength of a passing signal.

When these weights are modified, the data transferred through the network change and the network output alters. The neurons in a layer share the same input and output connections, but do not interconnect among themselves. All the nodes within a layer act synchronously. Hence, at any point of time, they will be at the same stage of processing. The activation levels of the hidden nodes are transmitted across connections with the nodes in the output layer. The level of activity generated at the output nodes is the network's solution to the problem presented at the input nodes. Each node multiplies every input by its weight, sums the product, and then passes the sum through a transfer function to produce its result. At the beginning of training the network weights are initialized, either with a set of random values or based on some previous experiences. The weights are optimized to get a specific response from an ANN. When these weights are modified, the data transfer through the ANN changes and the overall network performance alters. The learning algorithm adjusts the weights such that for a given input, the difference between the network output and the actual output is small.

In this paper, the generalized Delta rule is used to train a multi-layer perceptron for THM formation. As an output, the water quality is produced by presenting an input pattern to the network. According to the difference between the produced output and the observed, the parameters of network are adjusted to reduce the output error. The error at the output layer propagates backward to hidden layer, until it

reaches the input layer. Because of feedback propagation of error, the generalized Delta rule is also called by error back propagation algorithm. The output from node i, Oi, is connected to the input node j through the interconnection weight Wij. Unless node k is one of the input nodes, the state of node k is :

$$O_k = f(\sum W_{ik}\, O_i) \tag{3}$$

where, $f(x) = \frac{1}{(1+e^{-x})}$, called transfer function, and the sum is the total of all nodes in the adjacent layer. This transfer function is usually a steadily increasing S-shaped curve, called a sigmoid function. The transfer function also introduces a nonlinearity that further enhances the network's ability to model complex function. The sigmoid function is continuous, differentiable everywhere, and monotonically increasing. The output is always bounded between 0 and 1, and the input to the function can vary between plus or minus infinity.

Let the resulting target(output) state node be t. Thus, the error at the output node can be defined as

$$E = \frac{1}{2}\sum_k (t_k - O_k)^2 \tag{4}$$

where node k is the output node. The gradient descent algorithm adapts the weights according to the gradient error, i.e.,

$$\Delta W_{ij} \alpha \frac{E}{W_{ij}} = \frac{E}{O_j}\frac{O_j}{W_{ij}} \tag{5}$$

Specially, we define the error signal as

$$\delta_j = \frac{E}{O_j} \tag{6}$$

With some manipulation, we can get the following generalized Delta rule:

$$\Delta W_{ij} = \eta\, \delta_j\, O_i \tag{7}$$

where η is an adaptation gain. The δ_j is computed based on whether or not node j is in the output layer. If node j is one of the output nodes,

$$\delta_j = (t - O_j)\, O_j\, (1 - O_j) \tag{8}$$

If node j is not in the output layer,

$$\delta_j = (t - O_j)\, O_j \sum k\, \delta_k\, W_{kj} \tag{9}$$

In order to improve the convergence characteristics, we can introduce a momentum term with momentum gain α to Equation (7).

$$\Delta W_{ij}(n+1) = \eta\, \delta_j\, O_i + \alpha\, \Delta W_{ij}(n) \tag{10}$$

where n represents the iteration index. Once the neural network is trained, it produces very fast output for a given input data. It only requires a few multiplications and calculations of a transfer function.

3. DESIGN OF MODELS

3.1 Description of models architecture

Three techniques adopted in this study are applied to develop prediction models. The models are evaluated by comparing the results in terms of the accuracy, convenience, and ease of use. The difference among models is mainly in the input structure for the purpose of investigating its impact on the output accuracy and determining the most appropriate one for the case of the THM formation in water distribution systems.

Model I is consisted by seven independent variables which are used as inputs with the measured data. This model is expressed in the multiple nonlinear form as follows,

$$DBP = k \cdot T^a \cdot pH^b \cdot TOC^c \cdot (Cl_2)_0^d \cdot Br^e \cdot UV254^f \cdot t^g \quad (11)$$

where k, a, b, c, d, e, f, g denote regression coefficients ; T = temperature(°C) TOC = total organic carbon(mg/L), $(Cl_2)_0$, = initial concentration of chlorination(mg/L), Br = bromide level(mg/L), UV254 = UV light absorbance (mg/L), t = reaction time(hr).

Model II is similar to model I, but it is expressed by linear form of each term. Six independent variables are used as inputs and k', l, m, n, o, p, q denote regression coefficients.

$$DBP = k' + l \times T + m \times pH + n \times TOC + o \times (cl_2)_0 + p \times (UV254) + q \times t \quad (12)$$

The ANN is a computing paradigm that may have more than one mode. The feed forward neural networks with back-propagation learning algorithm are the most widely used neural networks. This study employs three-layer networks. The configuration of a neural network includes determining the number of hidden layers, the number of nodes in each of the hidden layers, and the connection weights.

The ANNs are trained with a set of input and known output pairs called training set. Many learning examples are repeatedly presented to a network, and the process is terminated when the difference is less than a specific value. The final weight matrix of the trained network represents its knowledge about the problem. Model III and model IV are constructed and applied to the prediction to THM formation in pipeline system. The system equations of models structure are expressed as follows;

Model III

$$DBP = ANN [T, pH, Cl_2, TOC, DOC, UV254, t] \quad (13)$$

Model IV

$$DBP = ANN[T, pH, Cl_2, UV254, t] \quad (14)$$

The architecture of model III and model IV can be seen in Fig. 1 and Fig. 2, respectively.

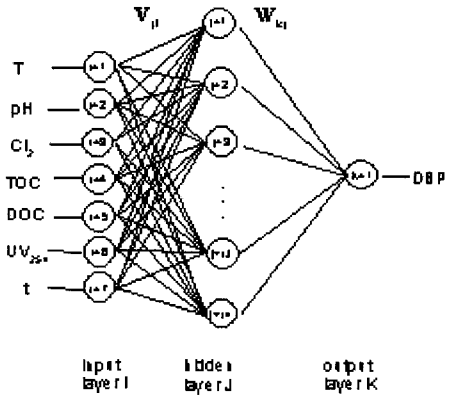

Fig. 1. Neural network model architecture of Model III

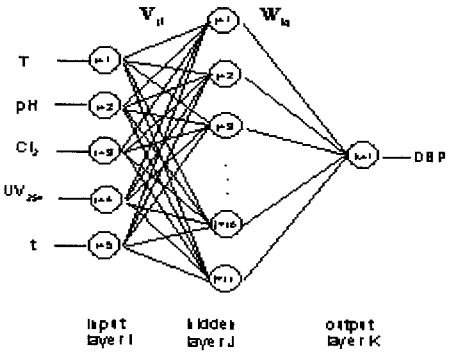

Fig. 2. Neural network model architecture of Model IV

3.2 Indicator of model performance

A determination coefficient(R^2) is one of the most commonly used performance measures for model evaluation. This provides information about model predictive capabilities. Equation for determination coefficient is given by :

$$R^2 = \sum_{i=1}^{N}(x_i - \hat{x}_i)^2 / \sum_{i=1}^{N}(x_i - \overline{x_i})^2 \qquad (15)$$

where $\overline{x_i}$ denotes the mean of measured value , x_i and $\hat{x}_i$ stands for the estimated value by each proposed model, and N is number of observations.

4. APPLICATION AND RESULTS

To apply the THM prediction modelling in pipeline system, water quality data are measured and collected. Data describing the pipeline system, which consist of 44 pipes and 45 nodes, and sampling sites characteristics are listed in Table 1. Ten sampling sites(W1~W10) are se-

Table 1. Characteristics of water pipeline system and samping sites

Uppr Node	Down Node	Daimeter (㎜)	Distance (m)	Discharge (CMD)	Velocity (m/sec)	Pipe No.	Travel time (hr)	Sampling sites
-	-	-	-	-	-	-	-	W1 (CJ Water Treatment Plant)
0	1	1650.	954.	247500.	2.8646	1	1978	
1	2	1200.	7126.	122500.	1.4178	2	1.5790	
2	3	1200.	626.	120500.	1.3947	3	1410	
3	4	1200.	2952.	119500.	1.3831	4	6705	W2 (G Water Tank)
3	46	150.	1360.	500.	0058	46	1.1536	
3	47	250.	1900.	500.	0058	47	4.4768	
4	5	1200.	400.	118500.	1.3715	5	0916	
5	6	1200.	148.	112200.	1.2986	6	0358	
5	49	350.	1910.	6300.	0729	49	7001	W3 (K Water Tank)
49	50	200.	850.	800.	0093	50	8011	
49	51	350.	2774.	5500.	0637	51	1.1646	
6	7	1100.	2456.	112200.	1.2986	7	4993	
7	8	2000.	695.	112200.	1.2986	8	4670	
8	9	1100.	3191.	112200.	1.2986	9	6487	
9	10	1100.	2751.	112000.	1.2963	10	5602	W4 (J Pump Station)
10	11	1100.	46.	112000.	1.2963	11	0094	
11	12	1100.	58.	112070.	1.2971	12	0118	
12	13	1100.	325.	112000.	1.2963	13	0662	
13	14	2000.	665.	112000.	1.2963	14	4477	
14	15	1100.	3232.	112000.	1.2963	15	6582	
15	16	1100.	633.	111000.	1.2847	16	1301	
16	17	1100.	2877.	110000.	1.2731	17	5965	
17	18	2000.	1260.	110000.	1.2731	18	8637	
18	19	1100.	1560.	110000.	1.2731	19	3235	
19	20	1100.	3874.	109500.	1.2674	20	8069	W5 (S Assembly House)
19	55	100.	1420.	500.	0058	55	5353	
20	21	1100.	854.	107600.	1.2454	21	1810	
20	60	300.	1693.	1900.	0220	60	1.5116	
60	61	300.	52.	1900.	0220	61	0464	W6 (M Pump Station)
61	62	300.	73.	1900.	0220	62	0652	
21	22	1100.	5100.	107600.	1.2454	22	1.0810	W7 (CA Water Treatment Plant)
22	23	1100.	26.	107600.	1.2454	23	0055	
23	24	700.	5266.	28300.	3275	24	1.7187	
24	67	250.	300.	2000.	0231	67	1767	W8 (D Industrial Zone)
24	25	700.	273.	26300.	3044	25	0959	
25	26	700.	2444.	25300.	2928	26	8922	
26	27	700.	100.	20300.	2350	27	0455	
27	28	700.	4184.	19000.	2199	28	2.0339	W9 (B Water Tank)
27	70	300.	1278.	1300.	0150	70	1.6678	
28	29	700.	1155.	16000.	1852	29	6667	
28	71	300.	3263.	1200.	0139	71	4.6130	
71	72	250.	420.	1200.	0139	72	4123	
72	73	300.	810.	1200.	0139	73	1.1451	
73	74	200.	1501.	1200.	0139	74	9431	W10 (D Pump Station)
28	80	250.	7600.	1800.	0208	80	4.9742	

Table 2. The correlation and determination coefficients of results for each model

Index	Correlation coefficient	Determination coefficient	Remarks
Model I	0.917	0.841	Multiple nonlinear regression
Model II	0.937	0.878	Multiple linear regression
Model III	0.986	0.972	ANN
Model IV	0.977	0.955	ANN

lected in the system where the measuring works of water quality data are available such as pump stations, water tank and public facilities etc. Also, the reaction time and hydraulic properties in Table 1 are calculated easily by using KYPIPE model.

Correlation analysis are performed using collected water quality data. Input variables are selected by the consideration of the results from correlation analysis. The linear relationships between THM formation and other water quality parameters(temperature, pH, Cl_2, TOC, DOC, UV254) can be shown in Fig. 3 through Fig. 9. From the analysis, it is found that THM formation is correlated in the order temperature, DOC, TOC, pH, Cl2, UV254. It is important factor to predict THM formation in pipeline system although the correlation coefficient of reaction time shows lowest value among water quality data as 0.006.

From the analysis of water quality data, the model equations of multi-regression methods are developed as follow;

$$DBP = 6.188 \times T^{0.4794} \times pH^{-2.869} \times Cl_2^{0.5107} \times UV254^{0.3598} \times t^{0.06916} \quad (16)$$

$$DBP = 0.05867 + 0.000687T - 0.009643pH + 0.01095Cl_2 + 0.4876UV254 + 0.0002916t \quad (17)$$

The correlation and determination coefficients of the results for each model are presented in Table 2 and the water pipeline system characteristics under consideration are shown in Table 2. In the prediction results of proposed four models model III shows the excellent prediction capability. The determination coefficient of model III is 0.972 as shown in Table 2.

The relationship between measured and predicted THM are shown in Fig. 10 through Fig. 13 to investigate the applicability and performance of models. The variation of learning error

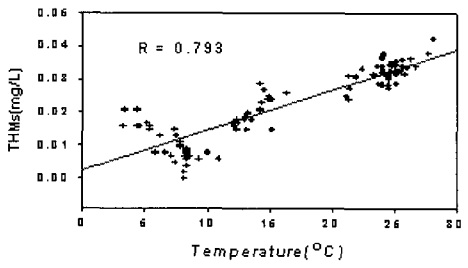

Fig. 3 Relationship between THMs and T

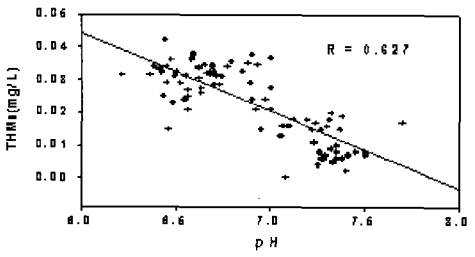

Fig 4. Relationship between THMs and pH

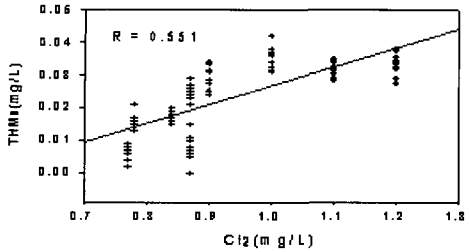

Fig. 5 Relationship between THMs and Cl_2

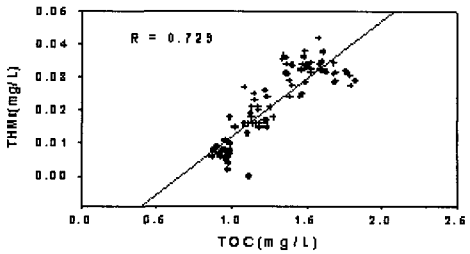

Fig. 6 Relationship between THMs and TOC

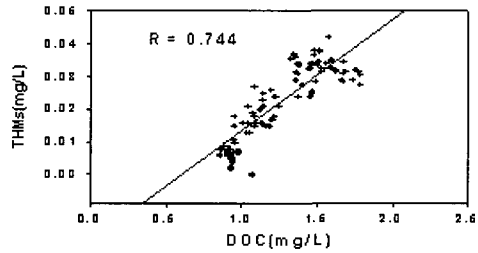

Fig. 7 Relationship between THMs and DOC

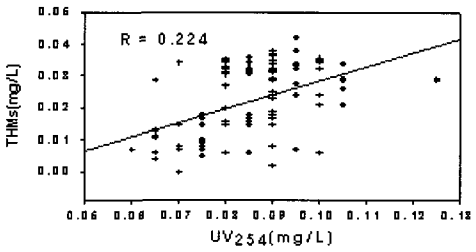

Fig. 8 Relationship between THMs and UV254

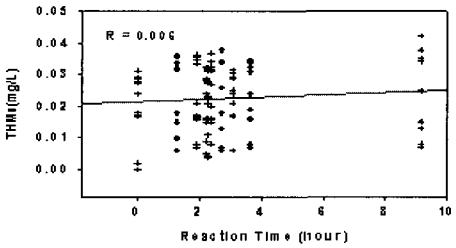

Fig. 9 Relationship between THMs and reaction time

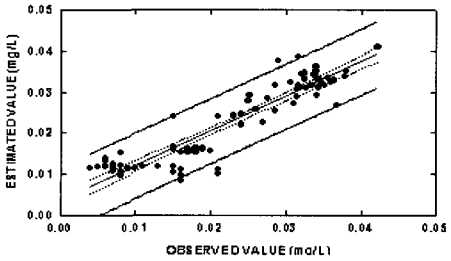

Fig. 10 Comparison of result by Model I

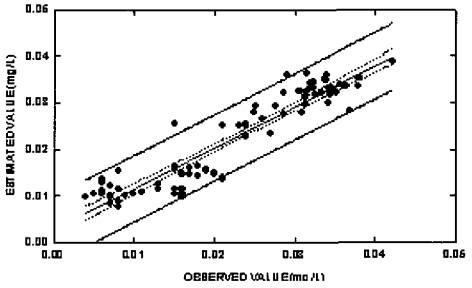

Fig. 11 Comparison of result by Model II

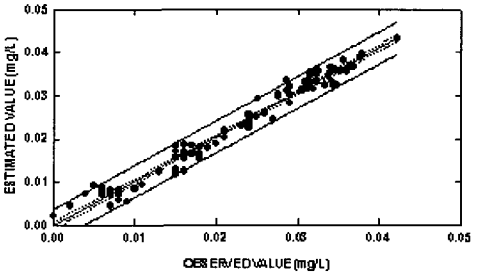

Fig. 12 Comparison of result by Model III

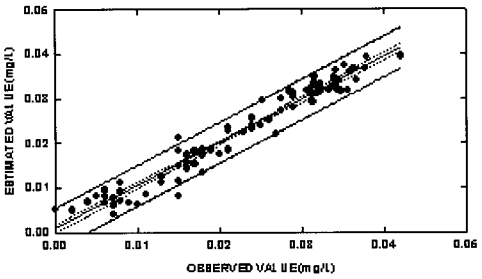

Fig. 13 Comparison of result by Model IV

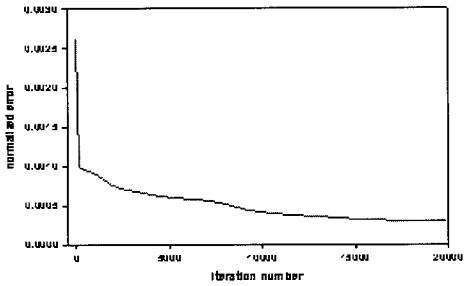

Fig. 14 Variation of learning error according to training iteration

of model III by iteration number can be seen in Fig. 14. It can be found that the learning error is continuously decreased as the training of the model is progressing.

5. CONCLUSIONS

Water quality prediction models in pipeline systems based on multiple linear/nonlinear and neural network are developed and presented. Especially, the relations between THM formation and other water quality parameters are shown and analyzed to formulate the process.

In the prediction results of proposed four models, model III based on neural network theory shows the excellent prediction capability. The ANN methodology has been reported to provide reasonably good solutions for circumstances where there are complex systems that may be poorly defined or understood using mathematical equations, problems that deal with uncertainty like water quality in pipeline systems. Especially, the reaction in pipeline system complicates defining and estimating the variation of water qualities. In this respect, neural network can be an effective and viable tool for not only the prediction of THMs formation but also other water quality factors in pipeline system.

Although the neural network method shows the usefulness in this study for modelling water quality in pipeline systems, the process of applications and data management (pre-processing and post-processing) are well-defined for the purposes of application areas. However, it has to be noted that further research is needed to fully understand its modeling capability.

REFERENCES

Amy G.L., Chadik Z.K. and Chowdhury Z.L. (1987). "A developing model for predicting trihalomethane formation potential kinetics," *J. of AWWA* 79(9), 89.

Bryant, E.A. and Fulton, G.P. and Budd, G.C. (1992). *"Disinfection alternatives for safe drinking water."* Van Nostrand Reinhold, New York.

Clark R.M. Smalley G., Godrich J., Tull R., Rossman L.A., Vasconcelos J.J and Boulus P.F. (1994). "Managing water quality in distribution systems : simulating TTHM and chlorine residual propagation," *J. Water SRT-AQUA* 43(4), pp. 182~191.

John Hertz, Andes Krogh, Richard G. Palmer, (1991). *"Introduction to the theory of neural computation."* Addison-Wesley Publishing

Company.

K. Urano et al. (1983). "Empirical rate equation for trihalomethans formation with chlorination of humic substances in water," Water Research, Vol. 12.

Kim J.H, and Kang K.W. Park, C.Y. (1992). "Nonlinear forecasting of streamflows by pattern recognition method," *Korean J. of Hydroscience,* Korean Ed., Vol. 25, No. 3.

Lisboa, P.G.J. (1992*). "Neural networks."* Chapman & hall, London, pp. 5~6.

Montgomery Watson, (1993). *"Mathematical modeling of the formation of THMs and HAAs in chlorinated natural water."* Final Report, AWWA.

Reckhow, D.A. Singer, P.C, and Malcom, R.L., (1990). "Chlorination og humic materials: Byproduct formation and chemical interpretation: Envir." Sci. & Technol., 24:11

Rossman L.A., Clark R.M., and Grayman W.M. (1994). "Modeling chlorine residuals in drinking water distribution systems," *J. Envir. Engrg.,* ASCE 120(4), pp. 803~829.

Singer, P. C., Barry, J. J. III., Palen, G.M. and Scrivner, A. E. (1981). *"Trihalomethane formation in north Carolina drinking waters."* AWWA, Aug.

Wood Don J. (1981). *"Computer analysis of flow in pipe networks including extended period simulation."* User's Manual, Office of engineering, Continuing education and extension, Univ. of Kentucky, Lexington, Ky.

Water Resources Research Institute, Korea Water Resources Corporation, Jeonmin-Dong 462-1, Yusung-Gu, Daejeon, Korea
(E-mail : juhwan@kowaco.or.kr)