MLP, ANFIS, and GRNN based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant

Chan Moon Kim and Manukid Parnichkun

ABSTRACT

Real-time determination of appropriate coagulant dosage under wide fluctuation of raw water quality in a water treatment plant (WTP) is a challenging task due to nonlinearity relation between coagulant dosage and raw water characteristics. In this research, three techniques, multilayer perceptron (MLP), adaptive neuro fuzzy inference system (ANFIS), and generalized regression neural network (GRNN), are applied to determine the coagulant dosage at Bansong drinking WTP. Each model is developed based on 8,760 historical data sets with hourly resolution for a whole year. Several statistical properties are determined to obtain the best-fit model from each method. The top performing models of each method are evaluated by external validation indices and absolute relative error according to nine turbidity zones. From the result, MLP and ANFIS models meet all conditions of validation indices, but GRNN cannot. The MLP shows the best result for high turbidity zones over 20 NTU as well as for overall performance. Meanwhile, ANFIS provides consistent results and better performance than MLP for low turbidity zones which have higher disorder of coagulant dosage data. The GRNN shows high accuracy for the highest turbidity zone which occurs during the rainy season. It is concluded that MLP, ANFIS, and GRNN can support operators effectively for real-time determination of coagulant dosage.

Key words | adaptive neuro fuzzy inference system, artificial neural networks, coagulant dosage, modeling, water treatment plant

Chan Moon Kim (corresponding author)
Manukid Parnichkun
School of Engineering and Technology,
Asian Institute of Technology,
P.O. Box 4,
Klong Luang, Pathumthani 12120,
Thailand
E-mail: *kcm0266@kwater.or.kr*

INTRODUCTION

Coagulation process is a vital process of a water treatment plant (WTP) since the process is intended to remove colloidal and fine particles by injecting chemical coagulant. Even though there have been significant advances of automation in WTP facilities, determination of coagulant dosage in the coagulation process is still challenging to automate due to the complexity of the coagulation process which has a nonlinear relation between many physical, chemical, and operational variables. Thus, the determination of coagulant dosage has been conducted traditionally by the operator's heuristic decision, or jar test which is a laboratory procedure. However, these practices are not able to respond to the abrupt changes in raw water quality (Joo *et al.* 2000) and are not adequate for real-time control (Yu *et al.* 2000). Since the water treatment process relies on the operator's experience and prior knowledge to control the coagulant dosages, it might result in excess or insufficient coagulant dosage caused by human error, particularly during the period of rapid variation of raw water characteristics. In this research, models which can support the operator's decision on coagulant dosage are developed and evaluated based on artificial neural network (ANN) and adaptive neuro fuzzy inference system (ANFIS).

A number of models used to determine coagulant dosage from raw water variables using ANN or ANFIS have been developed in the literature. Gagnon *et al.* (1997)

doi: 10.2166/aqua.2016.022

developed an annual model and four seasonal models for predicting the optimal alum dose using ANN for the Ste-Foy WTP in Quebec, Canada. Evans *et al.* (1998) applied ANFIS to predict coagulant dosage rate in Huntington WTP, UK, and showed that ANFIS was superior to a multiple regression model. Joo *et al.* (2000) developed a model for the Chungju WTP in South Korea, and showed that the performance of ANN was better than multi-variable regression. Deveughele & Do-quang (2004) developed online prediction of an optimal coagulant dosage using ANN in the WTP located at Viry in the vicinity of Paris. Larmrini *et al.* (2005) developed a soft sensor for online estimation of an optimal coagulant dosage using ANN on the basis of multilayer perceptron (MLP) for the Rocade WTP located in Marrakech, Morocco. Wu & Lo (2008) developed a multi-variant design model of predicting real-time coagulant dosage with MLP and ANFIS of grid partition type for the WTP in Taipei, Taiwan. Heddam *et al.* (2011) applied two different ANN techniques in Boudououa WTP, Algeria, and reported that a generalized regression neural network (GRNN) model had a consistently superior performance to the radial-basis function neural network model. Heddam *et al.* (2012) compared two ANFIS for modeling of coagulant dosage in drinking WTP of Boudouaou, Algeria, and showed that subtractive clustering model was more reliable than grid partition model. ANN models have been used for predicting coagulant dosage from both raw water and treated quality parameters. Zhang & Stanley (1999) developed a model that used turbidity of the treated water as an input for predicting alum dosage at the Rossdale WTP in Edmonton, AB, Canada. Yu *et al.* (2000) developed a model to predict the coagulant dosage using treated water turbidity at a WTP in Taipei City, Taiwan. Maier *et al.* (2004) developed a model to predict an optimal alum dosage by adding the turbidity and color of treated water as inputs based on jar tests in southern Australia. Griffiths & Andrews (2011) developed a model to predict an optimal alum dosage required to achieve the desired settled water turbidity at Elgin Area WTP, Canada.

Generally, it is necessary for an operator to determine the coagulant dosage in real time, especially when there is an unexpected rise of raw water turbidity. Thus, the role of the model to determine coagulant dosage becomes more important during this period. Even though many researchers have proposed various AI techniques to determine coagulant dosage, the accuracy of the models at high turbidity zone with sparse data density has not been investigated so far.

In this research, three different techniques, MLP, GRNN, ANFIS (subtractive clustering algorithm), are applied to determine coagulant dosage in drinking WTP. Model accuracy of the three techniques are then compared. Model evaluation is conducted under varying influent turbidity zones which contain not only stable but also steep rise conditions. This research is carried out under a full-scale scenario based on hourly historical data for an entire year. The data are extracted from SCADA system in Bansong WTP (Changwon, South Korea) in order to simulate the actual operation conditions as much as possible. The bench-scale data are generally not able to account for the simultaneous change in key process parameters, and always fail when applied to full-scale systems (Baxter *et al.* 1999).

METHODS

Materials

Bansong water treatment process

The data in this research are obtained from Bansong WTP in Changwon city, South Korea. The WTP is managed by Korea Water Resources Corporation (Kwater). The WTP has a water purification capacity of 120,000 m^3 per day and the intake of raw water is from Nakdong River. Figure 1

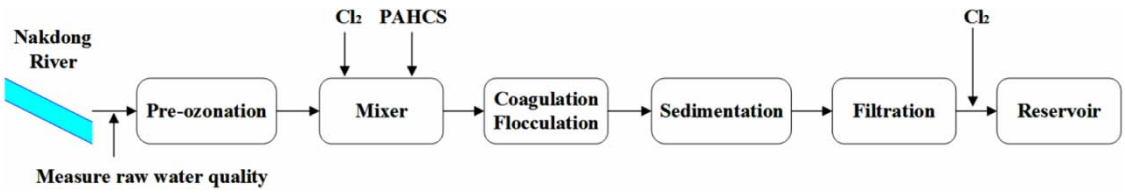

Figure 1 | Bansong WTP process showing chemical dose point and data collection point.

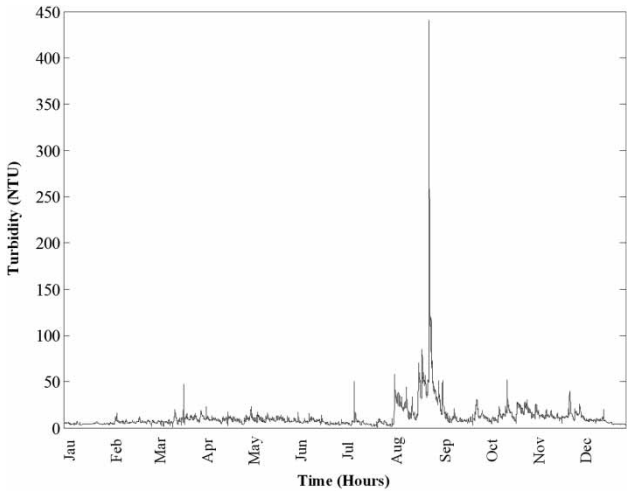

Figure 2 | Bansong WTP, influent hourly turbidity, 2014.

shows a schematic overview of the process at Bansong WTP. The treatment process consists of pre-ozonation, pre-chlorination, mixing, coagulation and flocculation, sedimentation, filtration, and post-chlorination. Bansong WTP has used polyaluminum hydrogen chloride silicate (PAHCS) as the coagulant since 2014.

Data collection and preprocessing

In Bansong WTP, turbidity, temperature, pH, and conductivity of the source raw waters are monitored by instruments in real time, and coagulant dosage is determined by operators based on the variation of the source water quality. In this research, hourly data are extracted during the entire period in 2014 from the database server to obtain input and output data required to develop the model. Thus, a total 8,760 records were gathered. In the collected data, there are some errors which show abrupt changes of the values, or no change in the values, or missing values. In this research, all data with errors are replaced by the value obtained from linear interpolation of the surrounding data.

Source data analysis

The raw water quality of Bansong WTP varies significantly during the year of 2014 because of considerable seasonal variation in the Nakdong River. Ninety percent of turbidity values of raw water are below 19.77 NTU. However, it increases to around 440 NTU with an extremely steep slope in the rainy season between August and September (Figure 2). The temperature shows considerable variation between 2.5°C and 30°C throughout the entire season. The PAHCS was injected at the rate of about 14 mg/L to 40 mg/L during the non-rainy season, while the highest rate with around 70 mg/L was dosed during the rainy season. The characteristics for the raw water quality and PAHCS dosage are summarized in Table 1.

Division of data and cross-validation

Cross-validation is a technique which is frequently applied in model development. It can be used to determine when to stop training for MLP and to compare generalization performance with other models. It needs three subsets of whole data: a training set, a validation set, and a testing set. Statistical properties such as mean and standard deviation of the three subsets must be close enough to guarantee that each subset represents the same statistical population of the domain (Maier *et al.* 2004). This research uses the data splitting method that was proposed by Baxter *et al.* (2001). All combinations of five data groups at the ratio of 3:1:1 are investigated to find the best data set retaining the closest statistical properties. As a result, the 8,760 data sets are divided into three subsets. Accordingly, the data sets are separated into 5,256, 1,752, and 1,752 data sets for the training,

Table 1 | The statistical properties of raw water quality and PAHCS dosage

Classification	Variable	Mean	Standard deviation	Min	Max	Median	90% percentile
Influent raw water	Turbidity (NTU)	11.46	16.55	0.49	440.85	8.00	19.77
	Temperature (°C)	16.43	8.09	2.52	29.90	17.66	25.72
	pH	7.67	0.35	6.76	8.85	7.63	8.19
	Conductivity (μs/cm^2)	299.70	89.55	110.20	537.17	291.93	415.72
Operational parameter	PAHCS dosage (mg/L)	29.64	6.76	14.36	69.60	30.08	35.74

validation, and testing, respectively. Table 2 gives the statistical properties of these data sets.

Techniques

MLP

ANN is a massive parallel information processing system resembling a biological nervous network of the human brain (Haykin 1998). MLP is the most commonly used ANN and has feed-forward hierarchical architecture and is generally used to map any input with corresponding output. MLP has been used for prediction and forecasting applications in many fields, and has also been applied successfully for coagulant dosage determination in the literature (Larmrini *et al.* 2005; Wu & Lo 2008). The back-propagation learning algorithm is the most widely used for optimizing connection weights in neural networks. Figure 3 shows a typical MLP.

ANFIS

ANFIS, which was introduced by Jang (1993), has also shown potential ability for coagulant dosage determination in the literature (Evans *et al.* 1998; Wu & Lo 2008). ANFIS combines the fuzzy inference system with multilayer feed-forward neural network forming a general structure where if-then rules with proper membership functions and the specified input–output pairs are used. Jang's ANFIS is generally represented by a five-layer feed-forward neural network. Figure 4 shows the ANFIS architecture corresponding to the Sugeno fuzzy model. In ANFIS, the parameters associated with the membership functions of input and output are trained using a hybrid learning algorithm that combines the least-squares estimator and the gradient descent method (Jang 1993). There are two most commonly used models for identification of fuzzy inference system in ANFIS. One is grid partition, the other is the clustering method. However, the grid partition method shows a critical drawback which causes a huge amount of calculation that results in inferior performance to clustering methods (Heddam *et al.* 2012). Thus, the rules are usually extracted from the observed data by using clustering techniques.

GRNN

The GRNN (Specht 1991) (Figure 5) based on non-linear regression is a neural network that can approximate any arbitrary continuous function by estimating a probability density function from training data. The main advantage of GRNN is that it does not require an iterative training procedure, thus it finishes training very rapidly. Moreover, it does present with the local minima problem like MLP and has consistent performance. When applying GRNN, it is important to find the optimal spread constant which is the smoothing parameter. In general, larger spread constant value leads to better generalization.

Table 2 | Statistical properties for training, validation, and testing sets

Data sets	Parameters	Mean	Standard deviation	Min	Max	Range
Training set	Turbidity	11.44	16.32	0.49	440.69	440.2
	Temperature	16.34	8.05	2.52	29.84	27.32
	pH	7.67	0.35	6.77	8.85	2.08
	Conductivity	300.67	90.4	110.51	536.76	426.25
	PAHCS dosage	29.64	6.75	14.36	66.7	52.34
Validation set	Turbidity	11.23	15.46	1.29	425.02	423.73
	Temperature	16.49	8.09	2.71	29.9	27.19
	pH	7.67	0.35	6.78	8.83	2.05
	Conductivity	300.85	89.1	110.2	536.05	425.85
	PAHCS dosage	29.64	6.75	15.09	66.63	51.54
Testing set	Turbidity	11.72	18.25	0.96	440.85	439.89
	Temperature	16.63	8.19	2.67	29.69	27.02
	pH	7.67	0.36	6.8	8.79	1.99
	Conductivity	295.62	87.4	11.17	537.17	526
	PAHCS dosage	29.66	6.79	15.17	69.6	54.43

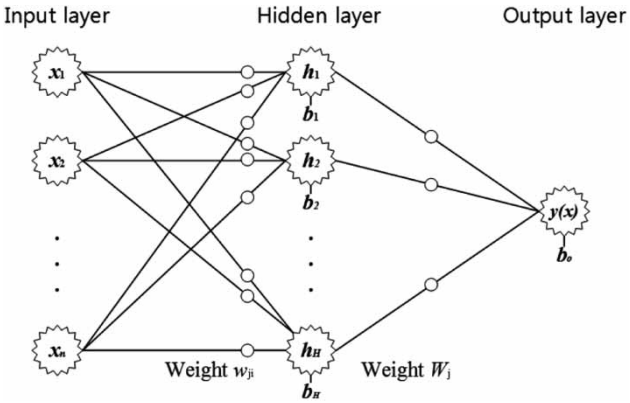

Figure 3 | MLP architecture.

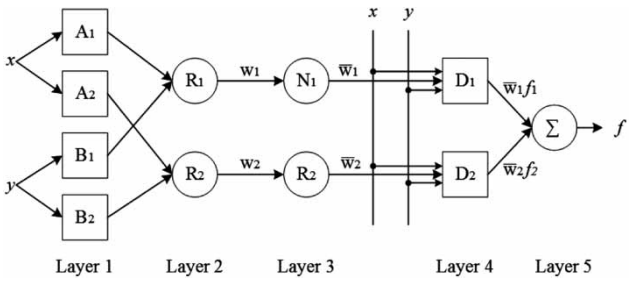

Figure 4 | Architecture of ANFIS.

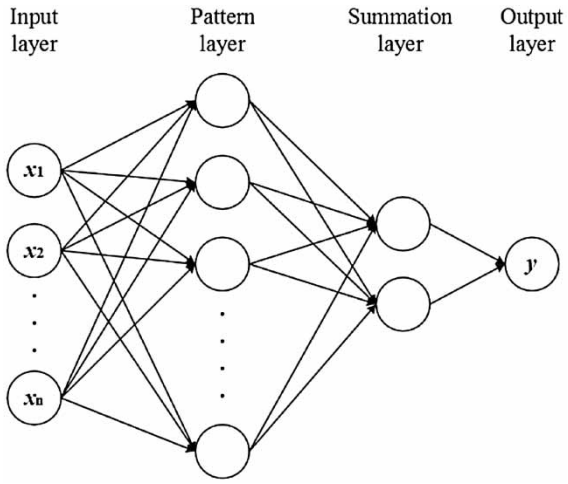

Figure 5 | Architecture of GRNN.

Model development

Input selection

The correlations among input parameters and PAHCS dosage are determined by a statistical index called Pearson's correlation coefficient. Table 3 shows the computed correlation coefficients between all parameters.

The correlation coefficient of turbidity and the PAHCS dosage is the highest at 0.534 compared with the other input, thus is the most relevant to the output. The coefficient magnitudes of pH, temperature, and conductivity vary between 0.135 and 0.247. Moreover, the input parameters are quite correlated among themselves as well. According to the result of Pearson analysis, eight possible combinations of inputs are used for developing the models of MLP, ANFIS, and GRNN (Table 4).

MLP and GRNN models

It was demonstrated that MLP with one hidden layer could approximate any continuous functions, given sufficient degrees of freedom (Hornik *et al.* 1989). Thus, one hidden layer is applied in this research. When using MLP with three-layer architecture, it is crucial to determine the number of neurons in the hidden layer. The ratio of the number of samples in the training data set to the number of connection weights as 10:1 was suggested in the literature in order to ensure good generalization of the model (Weigend *et al.* 1990). Therefore, the maximum number of neurons in the hidden layer can be determined from the number of training data samples (n_t) and the number of inputs (n_i):

$$n_h \leq \frac{n_t}{10(n_i + 1)} \tag{1}$$

Based on the previous researches, the evaluation of each model of MLP was performed by varying the number of hidden neurons from the initial size to the maximum size in Equation (1) in order to find the best model which has the lowest RMSE and the highest correlation coefficient for the test data set. A tangent sigmoid function and a linear function are applied for the hidden layer and output layer, respectively. The MLP model is trained by the Levenberg–Marquardt (LM) algorithm because it has good performance and fast learning speed with a simple structure (Hagan & Menhaj 1994). A cross-validation criterion is applied as a condition of stopping to prevent over-fitting. The parameters and

Table 3 | Pearson correlation coefficients of each input and PAHCS dosage

	Turbidity	Temperature	pH	Conductivity	PAHCS dosage
Turbidity	1.000	0.191	−0.381	−0.343	0.534
Temperature	0.191	1.000	−0.526	−0.541	0.153
pH			1.000	0.315	−0.247
Conductivity				1.000	−0.135
PAHCS dosage					1.000

Table 4 | Input combination for model development

Input type	Turbidity	Temperature	pH	Conductivity
1	I			
2	I	I		
3	I		I	
4	I			I
5	I	I	I	
6	I	I		I
7	I		I	I
8	I	I	I	I

their values used in MLP model development are summarized in Table 5.

The spread parameter constant for GRNN was investigated to find the optimal value by varying the range of the spread value from 0.1 to 5 with 0.1 resolutions. The spread value which results in best performance with the best RMSE and the highest correlation coefficient on the test data set is chosen. The MLP and GRNN models are built in the MATLAB R2014a (The Mathworks, Natick, MA, USA) environment by using the Neural Network Toolbox.

Table 5 | Parameter settings of MLP model development

	Parameter	Value
Geometry	Number of hidden layer	1
Transfer function	Hidden layer	Tangent sigmoid
	Output layer	Linear
LM algorithm	Performance function	Mean squared error
	Training strop criterion	Cross-validation stop
	Maximum continuous fail of validation	6
	Maximum epochs	500

ANFIS

Among various clustering methods, a subtractive clustering method is the best in conditions where the number of clusters for a given data set are not clearly determined (Talebizadeh & Moridnejad 2011). Hence, the subtractive clustering algorithm is applied in this research (Chiu 1994). In developing the ANFIS model, the membership functions are selected as Gaussian shape and the first order of the Sugeno fuzzy model is applied for the fuzzy inference system. In subtractive clustering, finding optimum influential radius is the most important (Chiu 1994). In general, a small radius makes many smaller clusters in the data space resulting in more rules. In this research, six radius values are applied to find the best candidate models: 0.3, 0.2, 0.1, 0.09, 0.08, and 0.07. The number of clusters is determined by the radius which results in the best performance for the test data set. The validation data set is used to cross-validate the fuzzy inference model to prevent over-fitting. The ANFIS models are trained in the MATLAB 2014a environment by using the Fuzzy Logic Toolbox. The parameter values used in development of the models are summarized in Table 6.

Models performance evaluation criteria

Normalized root mean square error (RMSE) is a descriptive index when the prediction performances are compared among models. Coefficient of correlation (R) is utilized to assess the strength of the relationship between the predicted and the observed values. Akaike's Information Criterion (AIC), which measures goodness-of-fit of the model to the data while penalizing model complexity caused by the number of variables in the model, can be used to select the optimal model (Qi & Zhang 2001). In order to

Table 6 | Parameter settings of ANFIS model development

	Parameter	Value
Geometry	Number of layers excepting	5
Membership function	Shape	Gaussian
	Number	No. of cluster by subtractive algorithm
Hybrid algorithm	Performance function	Root mean squared error
	Training strop criterion	Epoch number reach
	Avoid over-fitting	Minimum validation error criterion
	Maximum epochs	500

preliminarily evaluate the performance of all models with different inputs and parameters, these three statistical parameters are determined:

$$R = \frac{\sum_{k=1}^{n}(O_k - \overline{O_k})(P_k - \overline{P_k})}{\sqrt{\sum_{k=1}^{n}(O_k - \overline{O_k})^2 \sum_{k=1}^{n}(P_k - \overline{P_k})^2}} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n}(P_k - O_k)^2}{k}} \quad (3)$$

$$AIC = \log\left(\sqrt{\frac{\sum_{k=1}^{n}(P_k - O_k)^2}{k}}\right) + \frac{2m}{k} \quad (4)$$

where k is the number of data, P_k is the predicted value, and O_k is the observed value; $\overline{O_k}$ and $\overline{P_k}$ are the average values of the observed and the predicted values, respectively; m is the number of model inputs.

In this research, the external verification metrics are introduced to each best-fit model which has the best performance for MLP, ANFIS, and GRNN on the test data set. Kennedy *et al.* (2015) applied new criteria for coagulation process modeling. They evaluated the similarity between R^2 and $R_o^2/R_o'^2$ using the parameters of m and n. The values of m and n should have absolute value less than 0.1 if the model has good performance. They introduced the index R_m^2. This index evaluates the similarity between R_o^2 and R^2. The index value is greater than 0.5 when the model has a good performance. These metrics are defined as follows:

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(O_k - P_k)^2}{\sum_{k=1}^{n}(O_k - \overline{O_k})^2} \quad (5)$$

$$m = \frac{R^2 - R_o^2}{R^2} \quad (6)$$

$$n = \frac{R^2 - R_o'^2}{R^2} \quad (7)$$

$$R_m^2 = R^2\left(1 - \sqrt{|R^2 - R_o^2|}\right) \quad (8)$$

where R^2 is coefficient of determination which represents how much variation of the predicted value can be explained by the model. The model can be regarded as accurate if the R^2 is greater than 0.8. R_o^2 represents the coefficient of determination between the value from the perfect fit line and the observed value. $R_o'^2$ represents the coefficient of determination between the value from the perfect fit line and the predicted values.

Each best-fit model by MLP, ANFIS, and GRNN is evaluated using these external verification metrics and the absolute relative error percentage (ARE %).

$$ARE(\%) = \sum_{k=1}^{n}\left|\frac{P_k - O_k}{O_k}\right| \times 100 \quad (9)$$

Finally, the performances of the best-fit model are evaluated depending on various turbidity zones of raw water by mean absolute percentage error (MAPE), that is, mean value of ARE percentage. MAPE allows the user to notice percentage difference between the observed and the predicted values easily (Baxter *et al.* 2002).

RESULTS AND DISCUSSION

Each of the four candidate models from MLP, ANFIS, and GRNN methods with different input and architecture are selected by statistical evaluation on training, validation,

and testing data set, as shown in Table 7. In this table, the bold values denote the best-fit model of each method. It should be noted from the table that the models with all four input variables (input type 8) not only have the best-fit performances, but also show the most effective results with the smallest AIC values compared with the models with less variables. This is evidence of the fact that the four variables of raw water quality have a strong relationship with coagulant dosage. Moreover, it is proven that turbidity and temperature are the most relevant parameters for predicting PAHCS dosage. MLP and ANFIS models perform much better than the GRNN model. According to the testing results, the MLP(8,93) model is the best model considered with regard to RMSE and R.

Figure 6 gives the MLP performance of input type 8 with variation of the hidden neuron number from 3 to 105. As shown in the figure, the optimal number of hidden neurons is 93. The GRNN performance of input type 8 with variation of the spread constant is given in Figure 7. From this figure, it is noticed that spread constant of 1.0 is the best value for the

Table 7 | The performance of the candidate models from MLP, ANFIS and GRNN

	Training		Validation		Testing		
Model	**RMSE (mg/L)**	**R**	**RMSE (mg/L)**	**R**	**RMSE (mg/L)**	**R**	**AIC**
MLP(2,105)	3.105	0.888	3.244	0.877	3.164	0.885	2.306
MLP(5,86)	2.431	0.933	2.627	0.921	2.626	0.922	1.935
MLP(6,122)	2.397	0.935	2.556	0.925	2.476	0.931	1.816
MLP(8,93)	1.668	0.969	2.008	0.955	**1.959**	**0.957**	**1.350**
ANFIS(2, 0.8)	3.206	0.880	3.338	0.869	3.232	0.879	2.349
ANFIS(5, 0.7)	2.418	0.934	2.709	0.916	2.654	0.920	1.955
ANFIS(6, 0.8)	2.349	0.937	2.532	0.927	2.509	0.929	1.843
ANFIS(8,0.9)	1.766	0.965	2.046	0.953	**2.007**	**0.953**	**1.398**
GRNN(2,0.4)	2.736	0.915	3.229	0.878	3.160	0.885	2.303
GRNN(5,0.3)	2.257	0.943	3.016	0.895	2.948	0.901	2.166
GRNN(6,1.0)	1.602	0.972	2.615	0.922	2.560	0.926	1.883
GRNN(8,1.0)	1.547	0.974	2.572	0.925	**2.521**	**0.929**	**1.854**

Model definition: MLP(input type, number of hidden neuron); ANFIS(input type, cluster radius); GRNN(input type, spread constant).

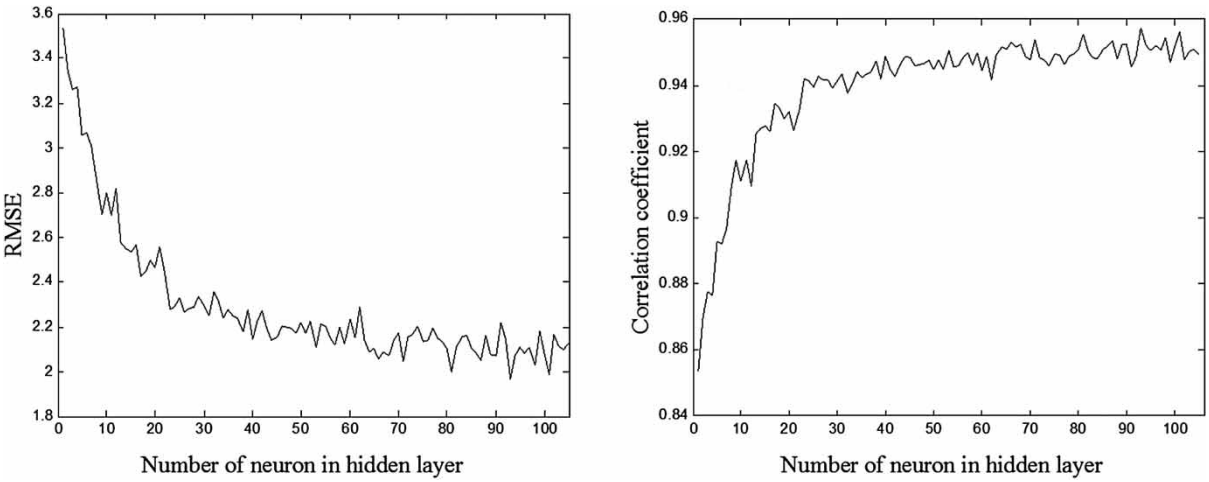

Figure 6 | Variation of the hidden neuron number and performance of the MLP model with input type 8.

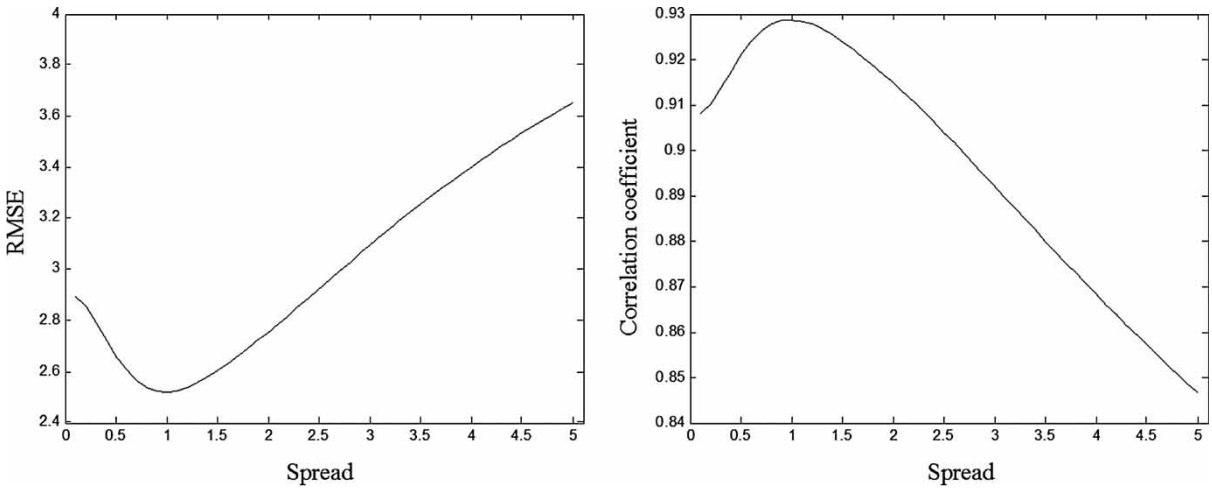

Figure 7 | Variation of the spread constant and performance of the GRNN model with input type 8.

model. A total 48 different models of the ANFIS with different input combination and the radius are simulated to find the best model. As shown in Table 7, the ANFIS(8,0.9) model is found to be the best model with 72 fuzzy rules.

According to the result of external verification indices in Table 8, the MLP and ANFIS models fulfill all the conditions of assessment, which means the prediction performances of the two models are very high and reliable. The GRNN model is not satisfied by m and n requirements, but it still meets the other two conditions. From the results of ARE evaluation, it is shown that the ANFIS model provides the most consistent prediction showing around 10% lower relative peak error than the MLP model. Figure 8 shows the time-series graphical plots of both the observed value and the best-fit value by MLP, ANFIS, and GRNN of the testing phase. It is clearly seen from the graphs that the predicted values of MLP and ANFIS are closer to the corresponding observed PAHCS dosage value than the value of the GRNN model, and the prediction value of ANFIS follows the observed value more smoothly.

The statistical evaluation indices used so far are global indices and do not present any information about the prediction accuracy on turbidity variation. Therefore, in order to test the robustness of each best-fit model with turbidity variation, the best-fit models of MLP, ANFIS, and GRNN according to nine turbidity zones of raw water are evaluated by MAPE criterion. This evaluation is based on 3,504 data from the validation and testing data sets, which are not used directly in training the models, in order to reach the general conclusion.

As given in Table 9, it is seen that the MLP model provides better results with about 5.2% of MAPE for the whole turbidity range, whereas GRNN is the most inferior model with around 6.5% of MAPE. However, it is obviously noticeable that each model is best-fit in certain turbidity zones. For turbidity less than 5 NTU, GRNN provides the best performance slightly over the ANFIS model. For turbidity from 5 NTU to 20 NTU which are the prevailing qualities of raw water, ANFIS is better than GRNN and MLP models. For high turbidity with a range from 20 NTU to 60 NTU, MLP results in better prediction ability than the others. For the highest turbidity zone over 60 NTU, which is critical for operation, GRNN shows better prediction than the other two models.

Table 8 | Statistical analysis of the best-fit models of MLP, ANFIS, and GRNN

Index	Criterion	MLP(8,93)	ANFIS(8,0.9)	GRNN(8,1.0)
m	$\|m\| < 0.1$	−0.0908	−0.0958	**−0.1596**
n	$\|n\| < 0.1$	−0.0903	−0.0952	**−0.1575**
R_m^2	$R_m^2 > 0.5$	0.6522	0.6428	0.5423
R^2	$R^2 > 0.8$	0.9167	0.9126	0.8621
ARE (%)	Min	0.0055	0.0019	0.0022
	Max	49.3423	39.6897	51.0661
	Mean	4.5843	4.4735	4.9704

Bold numbers indicate the recommended condition value.

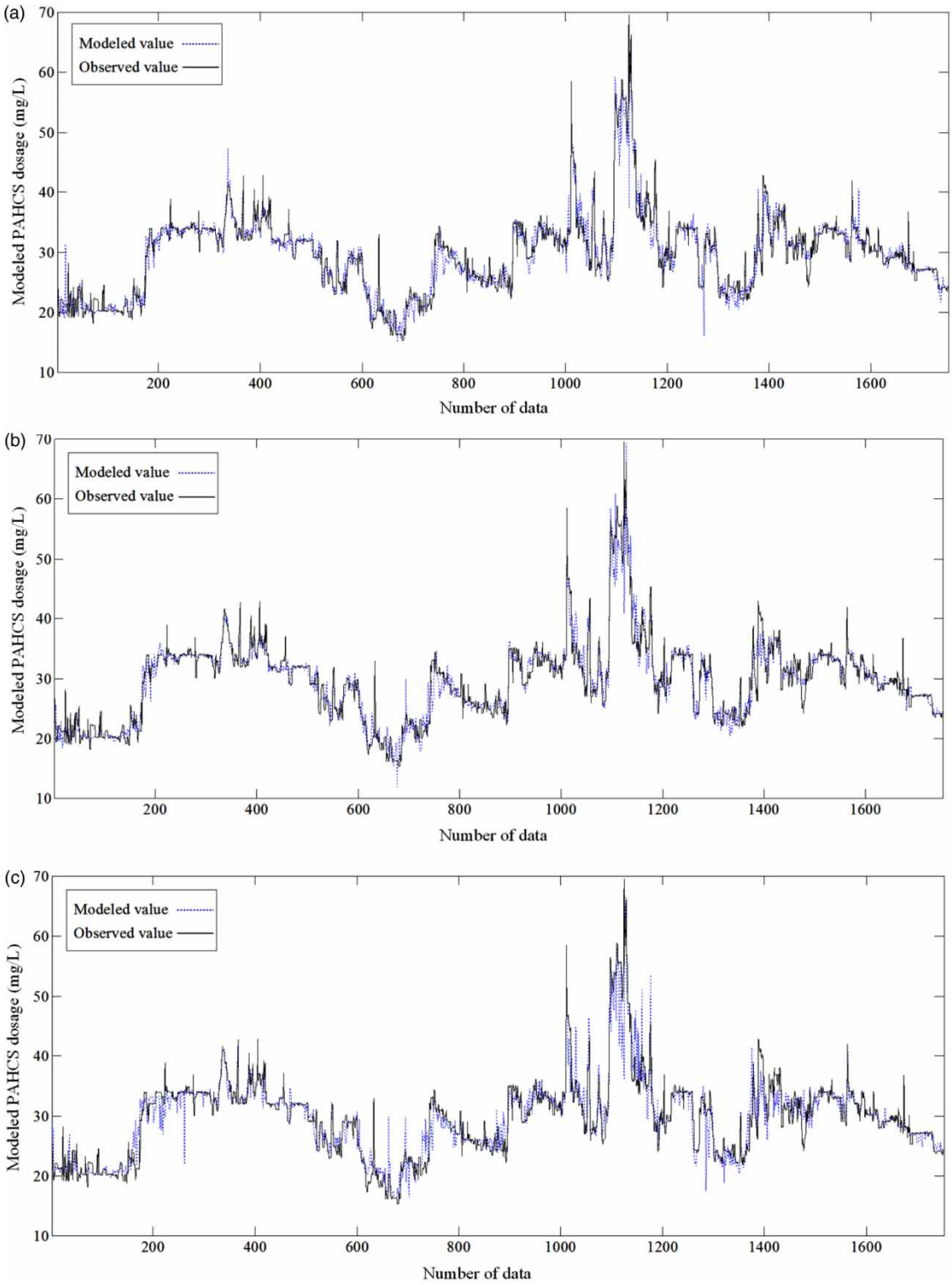

Figure 8 | The observed value and modeled PAHCS dosage by (a) MLP(8,93), (b) ANFIS(8,0.9), and (c) GRNN(8,1.0).

Table 9 | Comparing MAPE of MLP, ANFIS, and GRNN by raw water turbidity

Statistical measure	MAPE (%)			Ratio of average to standard deviation for PAHCS dosage	Number of data
Model	MLP(8,93)	ANFIS(8,0.9)	GRNN(8,1.0)		
Turbidity <5	4.028	3.685	**3.560**	5.307	666
5 <Turbidity <10	4.625	**4.306**	5.084	5.387	1,578
10 <Turbidity <15	4.652	**4.416**	5.180	6.359	705
15 <Turbidity <20	5.050	**4.901**	6.285	7.404	214
20 <Turbidity <30	**4.904**	6.034	5.179	9.927	171
30 <Turbidity <40	**5.733**	6.496	9.797	6.079	61
40 <Turbidity <50	**9.474**	10.854	14.404	4.699	43
50 <Turbidity <60	**4.195**	7.347	6.093	7.413	34
60 <Turbidity <450	3.851	6.390	**2.562**	8.108	32
Entire range of turbidity	**5.168**	6.048	6.460		

Bold numbers indicate the best performances in each turbidity zone.

Overall, the MLP and ANFIS models show promising prediction accuracy similarly, but MLP is better than ANFIS with respect to maintaining its accuracy at high turbidity zones. Meanwhile, as seen in Table 9, PAHCS dosage data show large disorder at low turbidity zones. Large disorder has a low value of the ratio of average to standard deviation (Mohammad & Mohammad 2008). It is noticed that ANFIS performs better than MLP as the degree of disorder of PAHCS dosage data increases with decreasing of the ratio value. This result confirms the results from the literature that neuro fuzzy system is good at handling large amounts of noisy data (Heddam *et al.* 2012). The GRNN model is inferior to MLP and ANFIS overall; however, it provides substantial improvement of prediction accuracy for sparsely extreme values of turbidity along with the advantage of fast training speed. Its accuracy increases up to 2.5 times of ANFIS and half the time of MLP. This result also confirms the fact that GRNN is preferable for sparse data in real-time situations (Heddam *et al.* 2011). Prediction results of three best-fit models with respect to the influent turbidity of over 60 NTU are graphically compared in Figure 9. It is confirmed that the value of the GRNN model is closer to the observed value than the other models on the whole.

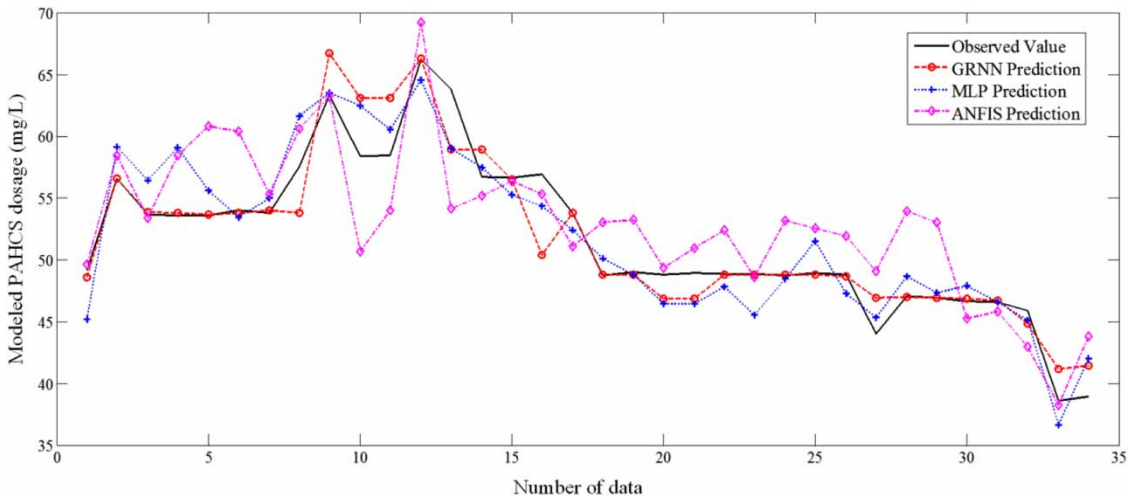

Figure 9 | Comparison of the performance of best-fit models for influent turbidity over 60 NTU.

Based on the strength of three techniques, performance improvement of many aspects can be achieved by combining each best-fit model in specific turbidity zones. Figure 10 shows the observed PAHCS dosage versus the value of the combined model which further improves the statistical results of prediction on the testing phase. The strength of each algorithm contained in the combined model can be summarized as follows; MLP has good prediction ability at high turbidity, ANFIS provides more consistent results and gives better prediction at low influent turbidity and high disorder data, and GRNN provides excellent performance for sparse data of extreme influent turbidity and swift training without local minima.

CONCLUSION

In this research, MLP, GRNN, and ANFIS with subtractive clustering techniques were employed to build models to predict coagulant dosage under full-scale conditions at a WTP. Various input combination and an extensive range of parameters of each algorithm were simulated to find the optimal architecture. Cross-validation technique was applied to ensure the generalization of the models. Effective statistical indices were used for evaluating the prediction performance of the models based on the data sets. The results of preliminary evaluation indicated that all the four candidate models obtained by MLP, ANFIS, and GRNN were capable of predicting PAHCS dosage with a high degree of accuracy. Especially, it was shown that turbidity and temperature were the most effective parameters for modeling. The accuracy of the models became higher when more variables were used as input, which proved that the coagulation process had complex nonlinearity characteristics of all the four parameters of influent water quality. By comparing the performance of the best-fit models from each method, it was seen that the MLP method led to the best result with low RMSE and high R while satisfying all external validation metrics. The MLP method had exceptional prediction ability at high turbidity zones over 20 NTU. However, it was found that MLP did not provide the best performance for all turbidity zones. ANFIS also provided good prediction ability similar to MLP, and it had the most consistent prediction results overall. In particular, it provided more reliable prediction accuracy than MLP at low turbidity zones which occupied the most data patterns for a year and inhered relatively high disorder of PAHCS dosage data. The GRNN model showed inferior performance compared to MLP and ANFIS models. It, however, provided outstanding prediction accuracy at the highest turbidity zone where operation was the most critical. From these results it is concluded that integration of the three techniques, depending on influent turbidity zones, can achieve more reliable and accurate prediction of coagulant dosage than individual use of a method during the rainy season as well as non-rainy season. In addition, GRNN can be utilized with sparse data for predicting coagulant dosage in the case of a real-time situation which requires a prompt decision on coagulant dosage.

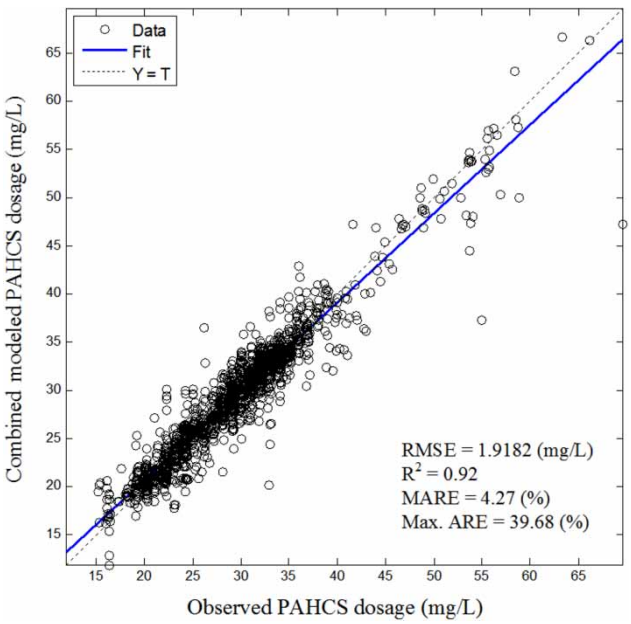

Figure 10 | Observed versus predicted PAHCS dosage using the combined model on the testing phase.

ACKNOWLEDGEMENT

The authors would like to thank the Kwater staff of Basnong water treatment plant in Changwon city, South Korea, for providing operational data and data collection support.

REFERENCES

Baxter, C. W., Stanley, S. J. & Zhang, Q. 1999 Development of a full-scale artificial neural network model for the removal of natural organic matter by enhanced coagulation. *Water Supply: Research and Technology-AQUA* **48** (4), 129–136.

Baxter, C. W., Tupas, R. R. T., Zhang, Q., Shariff, R., Stanley, S. J., Coffey, B. M. & Graff, K. G. 2001 *Artificial Intelligence Systems for Water Treatment Plant Optimization.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO, USA, p. 141.

Baxter, C. W., Stanley, S. J., Zhang, Q. & Smith, D. W. 2002 Development of artificial neural network models of water treatment processes: a guide for utilities. *Journal of Environmental Sciences* **1** (3), 201–211.

Chiu, S. L. 1994 Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems* **2** (3), 267–278.

Deveughele, S. & Do-Quang, Z. 2004 Neural networks: an efficient approach to predict on-line the optimal coagulant dose. *Water Science and Technology: Water Supply* **4** (5–6), 87–94.

Evans, J., Enoch, C., Johnson, M. & Williams, P. 1998 Intelligent based auto-coagulation control applied to a water treatment works. *Control '98. UKACC International Conference*, pp. 141–145.

Gagnon, C., Grandjean, B. P. A. & Thibault, J. 1997 Modelling of coagulant dosage in a water treatment plant. *Artificial Intelligence in Engineering* **11** (4), 401–404.

Griffiths, K. A. & Andrews, R. C. 2011 The application of artificial neural networks for the optimization of coagulant dosage. *Water Science and Technology: Water Supply* **11** (5), 605–611.

Hagan, M. T. & Menhaj, M. B. 1994 Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks* **5** (6), 989–993.

Haykin, S. 1998 *Neural Networks: A Comprehensive Foundation*, 2nd edn. Prentice-Hall, Upper Saddle River, NJ, USA, pp. 26–32.

Heddam, S., Bermad, A. & Dechemi, N. 2011 Applications of radial-basis function and generalized regression neural networks for modeling of coagulant dosage in a drinking water treatment plant: comparative study. *Journal of Environmental Engineering, ASCE* **137** (12), 1209–1214.

Heddam, S., Bermad, A. & Dechemi, N. 2012 ANFIS-based modeling for coagulant dosage in drinking water treatment plant: a case study. *Environmental Monitoring and Assessment* **184** (4), 1953–1971.

Hornik, K., Stinchcombe, M. & White, H. 1989 Multilayer feedforward networks are universal approximators. *Neural Networks* **2** (5), 359–366.

Jang, J. S. R. 1993 ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics* **23** (3), 665–685.

Joo, D. S., Choi, D. J. & Park, H. K. 2000 Determination of optimal coagulant dosing rate using an artificial neural network. *Water Supply: Research and Technology-AQUA* **49** (1), 49–55.

Kennedy, M. J., Gandomia, A. H. & Miller, C. M. 2015 Coagulation modeling using artificial neural networks to predict both turbidity and DOM-PARAFAC component removal. *Journal of Environmental Chemical Engineering* **3** (4), 2829–2838.

Larmrini, B., Benhammou, A., Le Lann, M. V. & Karama, A. 2005 A neural software sensor for online prediction of coagulant dosage in a drinking water treatment plant. *Transaction of the Institute of Measurement and Control* **27** (3), 195–213.

Maier, H. R., Morgan, N. & Chow, C. W. K. 2004 Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environmental Modelling & Software* **19** (5), 485–494.

Mohammad, Z. K. & Mohammad, T. 2008 Using adaptive neuro-fuzzy inference system for hydrological time series prediction. *Applied Soft Computing* **8** (2), 928–936.

Qi, M. & Zhang, G. P. 2001 An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research* **132** (3), 666–680.

Specht, D. F. 1991 A general regression neural network. *IEEE Transactions on Neural Networks* **2** (6), 568–576.

Talebizadeh, M. & Moridnejad, A. 2011 Uncertainty analysis for the forecast of lake level fluctuations using ensembles of ANN and ANFIS models. *Expert Systems with Applications* **38** (4), 4126–4135.

Weigend, A. S., Rumelhart, D. E. & Huberman, B. A. 1990 Predicting the future: a connectionist approach. *International Journal of Neural Systems* **1** (3), 193–209.

Wu, G. D. & Lo, S. L. 2008 Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Engineering Applications of Artificial Intelligence* **21** (8), 1189–1195.

Yu, R. F., Kang, S. F., Liaw, S. L. & Chen, M. C. 2000 Application of artificial neural network to control the coagulant dosing in water treatment plant. *Water Science and Technology* **42** (3–4), 403–408.

Zhang, Q. & Stanley, S. J. 1999 Real-time water treatment process control with artificial neural networks. *Journal of Environmental Engineering, ASCE* **125** (2), 153–160.

First received 22 March 2016; accepted in revised form 26 October 2016. Available online 8 December 2016