How Reliable Are ANN, ANFIS, and SVM Techniques for Predicting Longitudinal Dispersion Coefficient in Natural Rivers?

Roohollah Noori[1]; Zhiqiang Deng, M.ASCE[2]; Amin Kiaghadi[3]; and Fatemeh Torabi Kachoosangi[4]

Abstract: Determination of longitudinal dispersion coefficient (LDC) using artificial intelligence (AI) techniques can improve environmental management strategies for river systems. However, the uncertainty involved in AI models has rarely been reported. The main objective of this paper was to investigate the reliability of three AI-based techniques, including the artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), and support vector machine (SVM), for predicting the LDC in natural rivers. To that end, LDC predictions were first carried out using ANN, ANFIS, and SVM techniques. Then, a forward selection (FS) and gamma test (GT) were conducted to sort input variables according to their importance and effects on LDC prediction. Finally, uncertainties in the model predictions were analyzed to answer the question, "How reliable are ANN, ANFIS, and SVM techniques?" It was found that model inputs could not be satisfactorily sorted by a linear method (i.e., FS) due to the complex and nonlinear nature of LDC. Thus, the nonlinear GT technique was chosen as a suitable input selection method for prediction of LDC. The results or model input variables selected from the GT technique showed good consistency with previous researches. Furthermore, the reliability of ANN, ANFIS, and SVM models was calculated and tabulated by an uncertainty estimation for LDC prediction. A high uncertainty was found in the models although they predicted LDC appropriately. It was also found that the uncertainty in the SVM model was less than those in the ANN and ANFIS models for estimating the LDC in natural rivers. The ANFIS model performs better than the ANN model. **DOI: 10.1061/(ASCE)HY.1943-7900.0001062.**

Author keywords: Longitudinal dispersion coefficient; Artificial intelligence; Uncertainty analysis; Gamma test; Rivers.

Introduction

The advection-dispersion equation [Eq. (1)] is commonly used in one-dimensional (1D) water quality models, such as enhanced stream water quality model (QUAL2 E), river and stream water quality model (QUAL2 K), and water quality analysis simulation program (WASP), for prediction of water quality parameters in rivers and channels

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} = \mathrm{LDC}\frac{\partial^2 C}{\partial x^2} \tag{1}$$

where C = cross-sectionally averaged concentration; U = mean longitudinal velocity; t = time; LDC = longitudinal dispersion coefficient; and x = longitudinal coordinate along the direction of mean flow (Noori et al. 2011b). LDC has a great effect on pollutant transport in rivers. Therefore, prediction of LDC can be applicable to river water quality management and controlling strategies (Seo and Bake 2004). To determine LDC in rivers, many theoretical and empirical formulations, field experimental methods, and artificial intelligence (AI) techniques have been proposed.

The first theoretical method for predicting LDC was introduced by Taylor (1954). Elder (1959) applied this method to uniform flow. McQuivey and Keefer (1974) combined linear 1D flow with dispersion equations, which resulted in an LDC predictor equation. Other researchers preferred to develop simple methods based on a reasonable approximation of the triple integration, velocity deviation, and transverse turbulent diffusion coefficient proposed by Fisher (1975). Seo and Cheong (1998) applied a one-step regression method, proposed by Huber (1981), along with hydraulic and geometric data from 26 rivers in the United States for prediction of LDC. Kashefipour and Falconer (2002) also presented two regression models for predicting LDC in natural streams. The reported correlation coefficients of the two regression models were 0.84 and 0.80, respectively.

It should be noted that the LDC is generally controlled by a number of processes such as shearing advection, lateral mixing (Fischer et al. 1979; Rutherford 1994), transient storage (Bencala and Walters 1983), hypothetic exchange, and even effective diffusion in bottom sediment (Deng and Jung 2009; Deng et al. 2010). It is not easy to include all of the processes in a single method for estimation of LDC. As a result, there are significant differences between the so-called observed and predicted LDC values. It means that theoretical and empirical methods are not capable of predicting LDC accurately. Hence, some researchers attempted to determine LDC in natural rivers through field experiments. Fluorescent dyes, sulfur hexafluoride (SF6), and natural-artificial radiotracers are the most relevant soluble tracers (Graf 1995; Clark et al. 1996; Ho et al. 2006). Since these tracers are soluble in water, they are the best alternative for simulating the dispersion in surface waters (Smart and Laidlaw 1977). However, field experiments have faced some shortcomings and limitations. For example, many researchers have

[1]Assistant Professor, Dept. of Environmental Engineering, Graduate Faculty of Environment, Univ. of Tehran, 1417853111 Tehran, Iran (corresponding author). E-mail: noor@ut.ac.ir

[2]Associate Professor, Dept. of Civil and Environmental Engineering, Louisiana State Univ., Baton Rouge, LA 70803.

[3]Ph.D. Candidate, Dept. of Civil and Environmental Engineering, Univ. of Houston, Houston, TX 77204.

[4]Ph.D. Candidate, Dept. of Environmental Engineering, Graduate Faculty of Environment, Univ. of Tehran, 1417853111 Tehran, Iran.

Note. This manuscript was submitted on July 9, 2014; approved on June 1, 2015; published online on July 15, 2015. Discussion period open until December 15, 2015; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydraulic Engineering***, © ASCE, ISSN 0733-9429/04015039(8)/$25.00.

demonstrated that some tracers (e.g., chemical or fluorescent tracers) are nonconservative in natural waters (Smart and Laidlaw 1977). In addition, some toxic by-products are produced by the chemical reaction of some tracers in river water. They create different environmental problems for aquatic ecosystems (Ho et al. 2002, 2006). Furthermore, the high cost of soluble tracers is a limiting factor to their wide application (Clark et al. 1996).

Recently, data-driven approaches such as the artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), and support vector machine (SVM) have been widely used in the prediction of LDC. Tayfur and Singh (2005) trained and tested an ANN model to predict LDC in natural streams. Toprak and Savci (2007) conducted a comparison between a fuzzy model and seven existing predictor equations for LDC in natural channels. Results showed that outputs of the fuzzy model were more reliable than those of the predictor equation methods. Toprak and Cigizoglu (2008) computed LDC in natural streams using three ANN methods. Noori et al. (2009) presented two data-driven models, i.e., ANFIS and SVM, to predict LDC in natural rivers. In another work, genetic algorithms were used by Rajeev and Dutta (2009) to predict LDC in rivers. Riahi-Madvar et al. (2009) predicted LDC based on the ANFIS technique. They reported that the LDC values were accurately predicted by the ANFIS model. Noori et al. (2011b) trained and tested an ANN model with different training functions to achieve the best model for predicting the LDC in natural rivers. They showed that the ANN model with the Levenberg-Mardquate (LM) training function was superior. Azamathulla and Wu (2011) predicted the LDC in natural rivers and assessed the model performance using the coefficient of determination (R^2) and root-mean square error (RMSE). In another work, Azamathulla and Ghani (2011) proposed the genetic programming approach as a suitable method for predicting the LDC in natural streams.

Although the AI techniques showed promising performance in predicting LDC in natural rivers (Noori et al. 2009), some important questions, such as "How reliable are the predicted results of AI techniques?" or "How high is the uncertainty of AI techniques for predicting LDC in natural rivers?," remain to be answered. Generally, the total uncertainty involved in water quality modeling can be considered in both the model structure (Krzysztofowicz 2002; Montanari and Grossi 2008) and the model input data (Clark and Slater 2006; McMillan et al. 2011; Di Baldassarre and Montanari 2009). Unlike deterministic models which are commonly process-based, the structure of AI models, known as data-driven models, is a black-box approach. Therefore, tuning AI models is completely based on the data sampling patterns for calibration and verification of them. Thus, AI models are more sensitive to input parameters than deterministic models. As a result, AI models are subjected to more input-related uncertainty than deterministic ones.

While AI models have been widely used in various science and engineering fields, little is known about the uncertainty of them, particularly for the SVM model. Basically, uncertainty analysis of the SVM model has never been reported. This paper is intended to demonstrate a pioneering effort in the uncertainty analysis of AI models. Therefore, the uncertainty analysis of AI techniques (particularly the SVM model) is novel and useful to water quality management. Also, it can benefit emergency response managers and other decision makers for developing effective response strategies based on the results of the uncertainty analysis.

The objectives of this research are (1) to conduct data analysis using forward selection (FS) and gamma test (GT) techniques for selection of model input variables; (2) to develop ANN, ANFIS, and SVM models for LDC prediction; (3) to perform an uncertainty analysis of ANN, ANFIS, and SVM models; and (4) to compare the performance of ANN, ANFIS, and SVM techniques in terms of the prediction of LDC, based on their uncertainty analysis results.

Material and Methods

Experimental Data

Fluid properties, hydraulic characteristics, and geometric parameters are the main factors affecting LDC in natural streams (Deng et al. 2002). Therefore, the LDC model can be functionally expressed as Eq. (2)

$$\text{LDC} = f(U, H, B, u^*, \upsilon, S_f, S_n) \tag{2}$$

in which H = mean depth; B = width; S_f = bed shape factor; and S_n = sinuosity. It should be noted that S_f in natural rivers can be integrated in terms of shear velocity (u^*). Also, the effect of the Reynolds number, which is related to kinematic viscosity (υ), is negligible because of the completely turbulent flows in rivers (Seo and Cheong 1998). By considering these facts, Eq. (2) can be rewritten as

$$\text{LDC} = f(U, H, B, u^*, S_n) \tag{3}$$

In this research, modeling of LDC was carried out using 100 data sets measured in more than 30 streams in the United States. These data sets were collected from published works of Seo and Cheong (1998), Kashefipour and Falconer (2002), Tayfur (2006), Toprak and Cigizoglu (2008), and Etemad-Shahidi and Taghipour (2012). The statistical characteristics of the measured data sets are shown in Table 1.

GT Technique

GT, introduced by Agalbjorn et al. (1997), can be used to select input variables for a nonlinear and complex model in both noisy and low-noise situations (Durrant 2001). GT estimates the minimum mean square error (MSE) which contributes to input data selection. The selected data can be used in one of the nonlinear mathematical models. Suppose there is a set of data observations of the following form (Agalbjorn et al. 1997):

$$\{(\mathbf{x}_i, y_i),\ 1 \le i \le M\} \tag{4}$$

where the input vectors $\mathbf{x}_i \in R^m$ are m-dimensional vectors confined to a closed, bounded set $C \in R^m$ and the corresponding outputs $y_i \in R$ are scalars. The vectors $\mathbf{x}$ contain predicatively useful factors influencing the output y. The only requirement is the mean of the distribution of noise for the system to be zero, and the variance of the noise to be bounded. The noise estimated by the GT is actually a nonlinear equivalent of the sum squared error used in linear regression. The gamma statistic Γ is an estimate of the model's output variance that cannot be accounted for by a smooth data

Table 1. Statistical Characteristics of the Measured Data

Statistical characteristics	H (m)	B (m)	U(m/s)	S_n	u^* (m/s)	LDC (m^2/s)
Maximum	19.94	711.2	1.74	2.64	0.553	1,486.45
Minimum	0.22	11.89	0.13	1.07	0.02	1.9
Standard deviation	2.811	130.6	0.388	0.467	0.076	216.63
Mean	1.74	82.36	0.558	1.54	0.085	117.74
Median	0.91	40.54	0.45	1.32	0.069	41.4

model. The GT is based on $N[i,k]$, which are the kth $(1 \le k \le p)$ nearest neighbors $\mathbf{x}_{N[i,k]}$ for each vector $\mathbf{x}_i (1 \le i \le M)$. Specifically, the GT is derived from the delta function of the input vectors (Noori et al. 2011a)

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |\mathbf{x}_{N[i,k]} - \mathbf{x}_i|^2 \qquad (1 \le k \le p) \tag{5}$$

where $|\cdots|$ denotes Euclidean distance, and the corresponding gamma function of the output values (Agalbjorn et al. 1997)

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} |y_{N[i,k]} - y_i|^2 \qquad (1 \le k \le p) \tag{6}$$

where $y_{N[i,k]}$ = corresponding y-value for the kth nearest neighbor of $\mathbf{x}_i$ in Eq. (5). In order to compute Γ, a least-squares regression line is constructed for the p points $[\delta_M(k), \gamma_M(k)]$ (Noori et al. 2011a)

$$\gamma = A\delta + \Gamma \tag{7}$$

The intercept on the vertical axis ($\delta = 0$) is the Γ value, as can be shown (Durrant 2001)

$$\gamma_M(k) \to \mathrm{Var}(r) \text{ in probability as } \delta_M(k) \to 0 \tag{8}$$

where r = random variable which represents noise (Agalbjorn et al. 1997). Graphical depiction of the regression line in Eq. (7) provides useful information. First, it is noticeable that the vertical intercept Γ of the y-axis (or gamma axis) returns an estimate of the best achievable MSE using a modeling technique for unknown smooth functions of continuous variables (Evans and Jones 2002). Second, the gradient implies the model's complexity (a steeper gradient is an indication of higher model complexity). The GT results can be standardized by considering another term v-ratio, which returns a scaled invariant noise estimate between 0 and 1. The v-ratio is defined as

$$v\text{-ratio} = \frac{\Gamma}{\sigma^2(y)} \tag{9}$$

where $\sigma^2(y)$ = variance of output y, which enables researchers to judge how well the output can be modeled by a smooth function, regardless of the output range. A v-ratio close to 0 is an indication of a high degree of predictability of the given output y. If the v-ratio is close to 1, the output is equivalent to a random walk (Durrant 2001).

In practice, the GT can be achieved using the *winGamma* software (Durrant 2001). The authors believe that this technique is very effective and could be potentially used in many hydraulic nonlinear modeling efforts.

Artificial Intelligence Techniques

ANN

ANN as a "black-box" technique is an appropriate solution for modeling nonlinear and complex phenomena. Among various types of ANN models, the feed-forward neural network (FFNN) with back-propagation learning algorithm has been known as a universal predictor (Hornik 1989). FFNN's structure consists of (1) a training algorithm, (2) hidden layers, (3) the number of nodes in hidden layers, and (4) activation functions. More details about the ANN model can be found in Haykin (1994). In this research, FFNN with one hidden layer was used. Also, the LM algorithm was selected to optimize network parameters. Besides, tangent sigmoid and purelin transfer functions were chosen for hidden and output layers, respectively.

ANFIS

ANFIS modeling refers to the method involving the application of various learning techniques developed in the neural network literature to a fuzzy inference system (FIS) (Brown and Harris 1994; Nayak et al. 2004). For optimization of FIS in the ANFIS structure, FFNN is usually used. Contrary to the ANN, the ANFIS model utilizes a hybrid algorithm consisting of gradient descent and least-square methods to learn ANFIS from the data being modeled (Jang et al. 1997; Nayak et al. 2004). In this research, a first-order Sugeno FIS is used. Interested individuals are referred to Jang (1993) for more details about ANFIS, FIS structure, and the hybrid learning algorithm.

SVM

Generally, there are two types of SVM regression models: ε-SVM and ϑ-SVM models. In this research, ε-SVM regression was used. In the ε-SVM regression model, the deterministic function $f(\mathbf{x}) = \mathbf{w}^T \cdot \phi(\mathbf{x}) + b$ is employed in which $\mathbf{w}$ (vector of coefficients) and b (constant) are the regression function parameters and ϕ is the kernel function (Noori et al. 2011a). To find the regression function parameters, a sequential optimization process must be carried out. More details about the SVM model can be found in Vapnik (1998) and Abe (2005).

Determining the Uncertainty in AI Techniques

The AI techniques are black-box or semi–black box models. It causes complexity in the uncertainty analysis of AI techniques. In these models, input data are used to optimize the error function. In other words, the error function is directly related to the model's input data. The error function will be optimized with direct regard to the model's input data. It is, therefore, necessary to evaluate the model's performance through different calibration patterns in order to assess the model's uncertainty caused by changes in input data. To that end, the calibration pattern must be regenerated and outputs should be consequently computed by the model. This process is repeated many times. In this study, AI models were calibrated by a percentage of data. Data sampling process iteration was set to be 1,000 times. As a result, AI model parameters were reproduced 1,000 times by the patterns calibrated for each random sample. Therefore, a range of outputs related to the uncertainty of the 1,000 calibrated AI models would be determined by using this massive computational technique. To evaluate the AI models' uncertainty, the percentage of measured data bracketed by 95 percent of predicted uncertainties (95PPU) can be used. The 95PPU is computed by 2.5% (X_L) and 97.5% (X_U) of normal empirical distribution obtained from the 1,000 times forecasting process as follows (Aqil et al. 2007):

$$\text{Bracketed by 95 PPU} = \frac{1}{n}\mathrm{count}(Q|X_L \le Q \le X_U) \times 100 \tag{10}$$

In order to evaluate the average width of the confidence interval band, a d-factor parameter has been proposed as Eq. (11) (Abbaspour et al. 2007)

$$d\text{-factor} = \frac{\bar{d}_X}{\sigma_X} \tag{11}$$

where σ_X = standard deviation of the measured variable X; and $\bar{d}_X$ = average distance between the upper and the lower bands obtained from Eq. (12) (Abbaspour et al. 2007; Aqil et al. 2007)

$$\bar{d}_X = \frac{1}{k}\sum_{l=1}^{k}(X_U - X_L) \quad (12)$$

where k = number of observed data points.

Results and Discussion

Data Analysis

Two methods, including FS and GT, were employed for analyzing data sets and sorting the model input variables for LDC. The FS is a linear input determination technique which acts based on a simple correlation between input and output data (Chen et al. 2004). Therefore, the correlation between individual input variables and the LDC was determined. Results indicated that U had the highest correlation with LDC (with $R^2 = 0.41$). Therefore, it was selected as the most important input parameter to LDC. Thereafter, the remaining candidates, i.e., B, H, u^*, and S_n, were incorporated into the model. Results from the FS method showed that only two candidates were selected as input variables, including U and u^*. Thus, it was recommended to omit other candidate variables (B, H, and S_n). In the next step, the GT technique was utilized to select the best feature for LDC modeling. For selecting the best subset using the GT technique, different combinations of input data were explored. Results for GT application on the data set have been illustrated in Table 2. The table indicates that U is the most important variable because of the highest value for the noise variance (Γ) and smallest v-ratio, after its omission. Dispersion is the result of shear advection in space (Chapra 1997). Therefore, based on the definition of dispersion it is clear that U should be the most important variable affecting LDC in the longitudinal direction. Other important variables are B, H, u^*, and S_n, respectively. Generally, previous studies (Seo and Cheong 1998; Toprak and Cigizoglu 2008) demonstrated that the width-to-depth ratio plays an important role in LDC variations. Also, it is clear from Table 1 that B commonly varies in a wide range as compared with the depth H. Therefore, the effect of width-to-depth ratio on LDC is primarily due to width variations compared to depth variations. Besides, Table 2 indicates that u^* is the third most important parameter which can affect LDC in natural streams. It is noted that the influence of the S_f parameter is taken into account in u^*. Therefore, u^* can be a representative of itself and S_f (Seo and Cheong 1998). Table 2 indicates that the fifth most important parameter which can affect LDC in natural streams is S_n. Most of the studies, which utilized regression and AI techniques for LDC prediction, selected only U, B, H, and u^* as model input parameters. The main reason is that other parameters, such as S_n and S_f, have less effect on LDC. The effect of channel meandering, S_n, on LDC can be integrated into the shear velocity u^* (Seo and Cheong 1998). Therefore, the parameter S_n is less important to LDC compared to u^*.

By comparing results from FS and GT methods, it is found that the results from the FS procedure could be ignored because of three main reasons. First, previous studies (Fischer 1975; Liu 1977; Fischer et al. 1979; Koussis and Rodriguez-Mirasol 1998; Seo and Cheong 1998; Kashefipour and Falconer 2002; Seo and Baek 2004; Tayfur and Singh 2005; Toprak and Cigizoglu 2008; Noori et al. 2009) just selected the four variables (B, H, u^*, and U) as the main input parameters for LDC modeling. Second, results from dimensional analysis confirmed the significance of these four variables in LDC modeling (Seo and Cheong 1998). Third, a linear method like FS may not be able to analyze the complex and nonlinear nature of LDC. Therefore, GT was selected in this paper as the primary method for the selection of model input variables.

Table 2. GT Results on the Input Variables

Input variables	Gamma value (Γ)	v-ratio
All inputs	−0.31866	−1.27630
All inputs—B	−0.23053	−0.96212
All inputs—H	−0.24239	−0.92957
All inputs—U	−0.22361	−1.09450
All inputs—u^*	−0.24650	−0.90600
All inputs—S_n	−0.26012	−0.88956

ANN Model Development

In order to develop an ANN model, the input and output data were mapped to the domain $[-1,1]$. To avoid an overfitting problem, the stop training algorithm (STA) was used (Coulibaly et al. 2000). Due to the use of STA, the data were divided into a training set (60 data sets), validating set (20 data sets), and testing set (20 data sets). Training and validating data sets are generally known as the calibration set. Also, to compute the number of hidden neurons, an initial random number was employed. Afterward, the optimum number of hidden neurons was found to be 4 by a trial and error procedure. Results of the best-trained ANN model are presented in Table 3. In addition, Fig. 1 shows the results of model calibration and testing stages against observed data from the best-trained ANN model. The figure indicates that the predicted values of LDC generally have a good agreement with the observed data. On the other hand, it shows that the extreme LDC values obtained from the ANN model do not correspond to the observed ones. There was a significant difference between the predicted and observed extreme values. Therefore, although the ANN model generally produces acceptable performance in predicting LDC in natural rivers, it is not capable of predicting the extreme values accurately.

ANFIS Model Development

To run a fuzzy model two alternatives are available, including subtractive fuzzy clustering (requiring less computational effort) and grid partitioning (requiring more computational effort). To determine the uncertainty in the ANFIS model, it should be calibrated 1,000 times. In this work, the subtractive fuzzy clustering alternative was employed to reduce the computational times. Also, the optimal values for the search radius (r_a) were identified using a trial and error procedure by varying the r_a from 0.1 to 0.9 (in the increment of 0.05), leading to 0.36. Also, the optimum number of rules for the model was found to be 2. After the initial FIS structure was established, calibrations were carried out. In addition, the STA method similar to ANN model development was employed here to avoid the overfitting problem. The calibration and testing results of the ANFIS model are presented in Table 3. According to

Table 3. Calibration and Testing Results of ANN, ANFIS, and SVM Models

	ANN		ANFIS		SVM	
Statistic	Calibration	Testing	Calibration	Testing	Calibration	Testing
d-factor	1.51	1.48	1.12	0.95	0.78	0.82
Bracketed by 95PPU (%)	98	100	94	100	0.92	0.85
R^2	0.88	0.86	0.85	0.87	0.93	0.94
AARE (%)	1.5	2.1	1.3	1.5	1.1	1.2

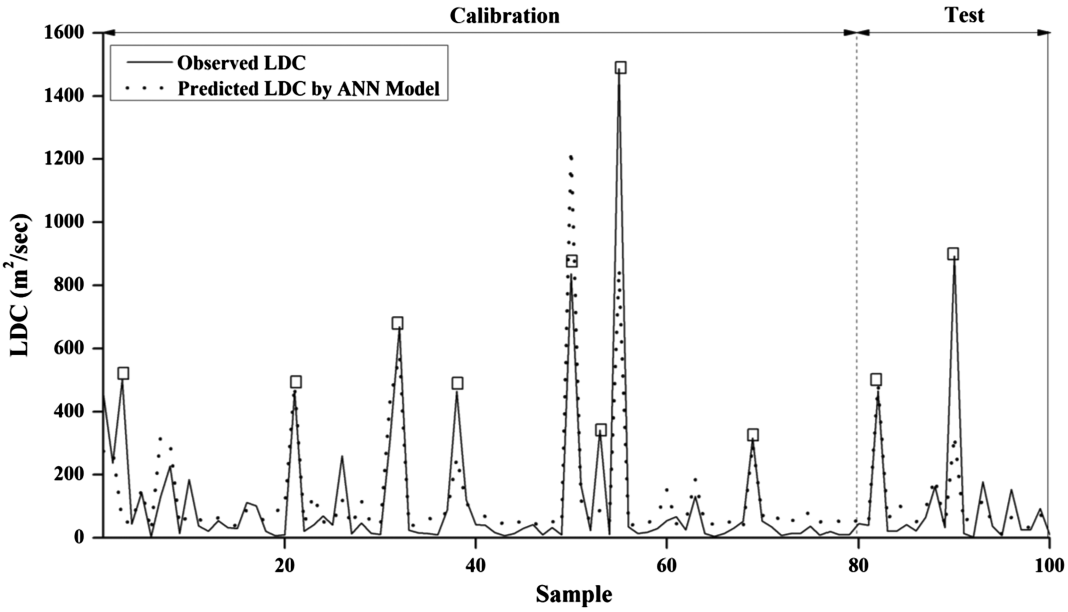

Fig. 1. Predicted and observed LDC for calibration and testing steps of the ANN model (square symbol □ indicates the extreme high values resulting from the model)

the table, although R^2 values are relatively similar in the testing step of ANN and ANFIS models, the ANFIS model has fewer errors in the prediction due to the better average of absolute relative error (AARE) as compared to the ANN model. In addition, the ANFIS model has better performance than the ANN model in predicting the extreme values of LDC, as evidenced in Fig. 2.

SVM Model Development

The radial base function (RBF) is used as the kernel function of the SVM model. Developing the SVM regression model, using the RBF kernel function, requires values for C and ε (tuning parameters of the SVM model) and the γ parameter. In this research, for fine-tuning these variables, a grid search algorithm with a tenfold cross-validation is performed. To implement the cross-validation technique, the data were divided into a training set (80 data sets) and testing set (20 data sets). The optimal parameters in the SVM regression model are achieved as $(C, \varepsilon, \gamma) = (15.50, 0.037, 2.46)$. Also, the training and testing results of the SVM regression model for LDC estimation are shown in Table 3 and Fig. 3. The figure indicates that the SVM model has better performance than the ANFIS and especially ANN, in predicting the extreme values of LDC.

Uncertainty Analysis of ANN, ANFIS, and SVM

One of the main factors for uncertainty analysis is the existence of a data set to make the random sampling process possible for the

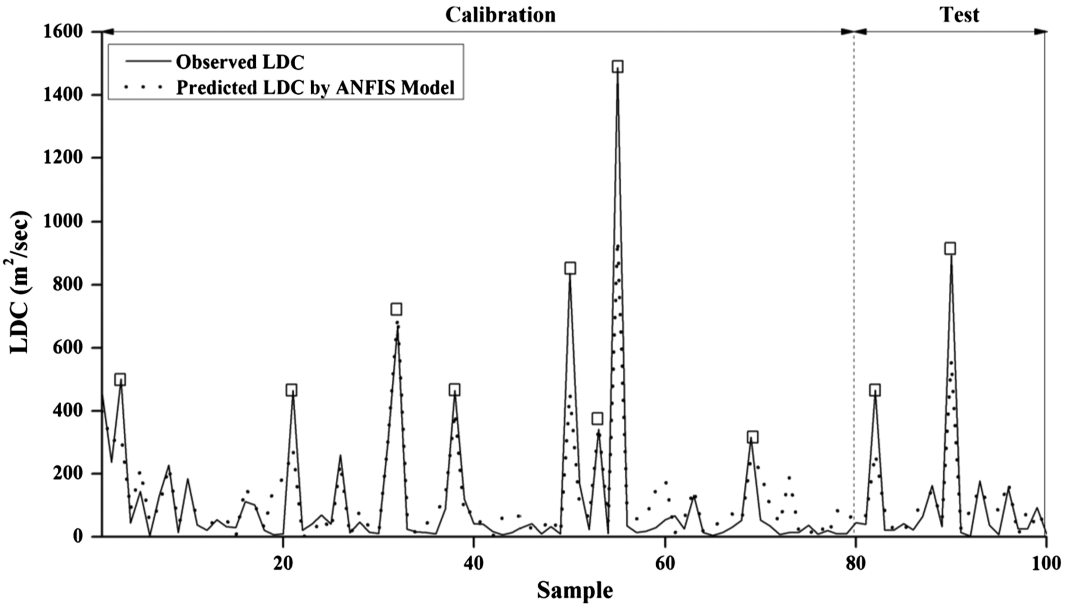

Fig. 2. Predicted and observed LDC for calibration and testing steps of the ANFIS model (square symbol □ indicates the extreme high values resulting from the model)

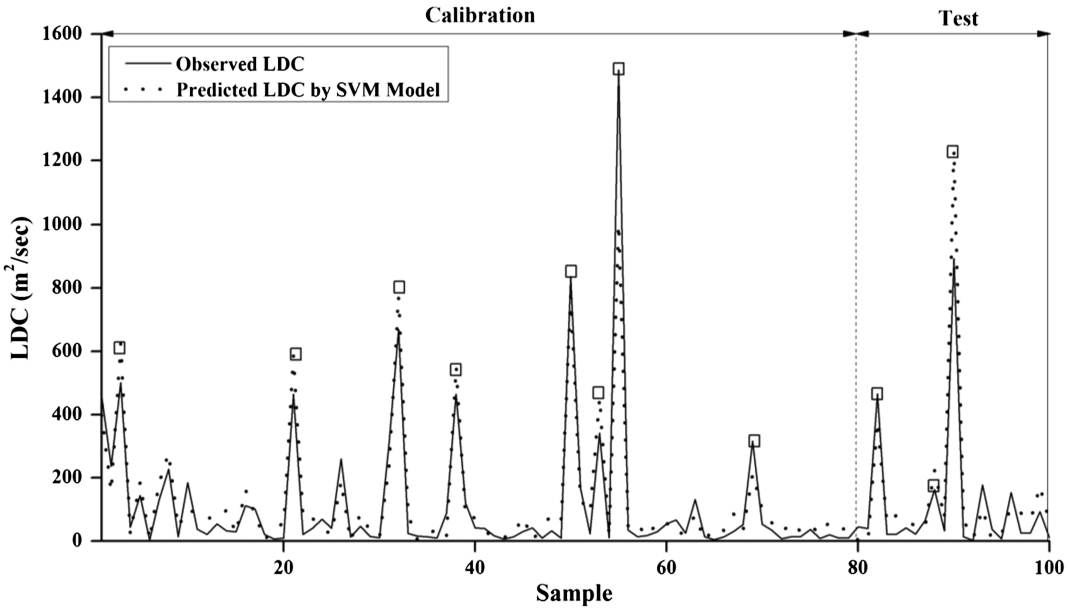

Fig. 3. Predicted and observed LDC for calibration and testing steps of the SVM model (square symbol □ indicates the extreme high values resulting from the model)

establishment of a training pattern. As mentioned previously, the available data sets were divided into three groups for developing the ANN and ANFIS models. In other words, validation data sets can be considered as a suitable data set to make the random sampling process of the training pattern possible. However, the available data were divided into two groups for developing the SVM model. An alternative group including 20 data sets was used in this paper for sampling the training in addition to the training and testing ones. Due to having only 100 data sets, there was an obligation for selecting the alternative group from the 100 available ones. This caused the authors to be left with only 80 data sets for training and testing the SVM model (60 data for training and 20 for testing). Therefore, a sampling algorithm was designed to provide the possibility of random replacement of 20% of the calibrating data sets with data sets in the alternative group. Finally, 1,000 times of sampling iteration resulted in 1,000 different modeling results for LDC prediction from each of the three models (ANN, ANFIS, and SVM). In the next step, 95PPU is determined by finding the 2.5th and 97.5th percentiles of the continuous distribution of the simulation results. The plots of 95PPU for the estimation of LDC during the calibrating and testing steps for the three models (ANN, ANFIS, and SVM) are shown in Figs. 4–6, respectively. Also, the *d*-factor and values bracketed by 95PPU in calibration and testing steps for AI models are listed in Table 3.

Reliability of Predicted Results for LDC

As mentioned earlier, significantly different LDC values were predicted using the AI technique-based models developed with almost

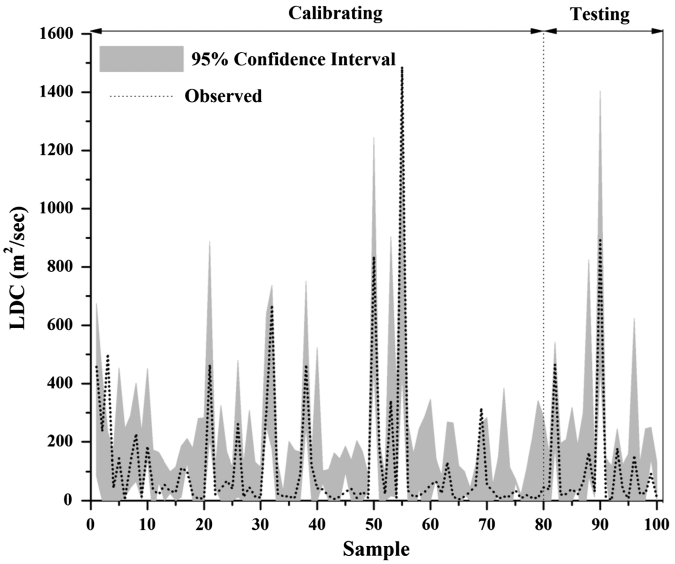

Fig. 4. 95% confidence intervals for the estimated LDC during calibrating and testing steps using the ANN model

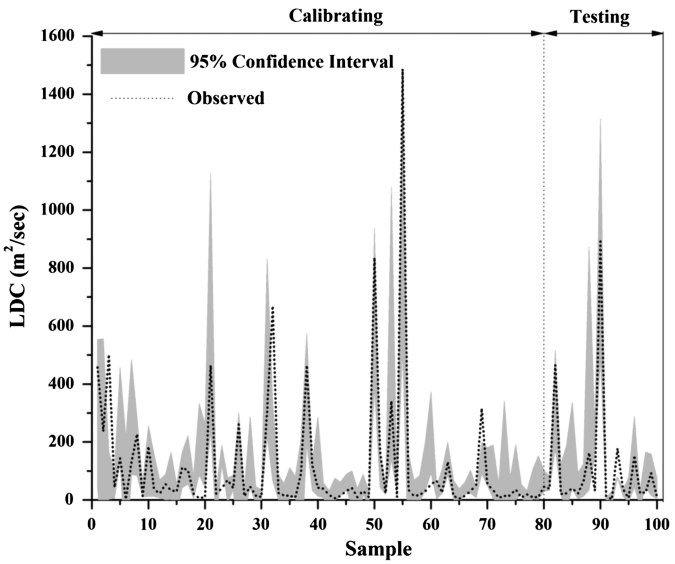

Fig. 5. 95% confidence intervals for the estimated LDC during calibrating and testing steps using the ANFIS model

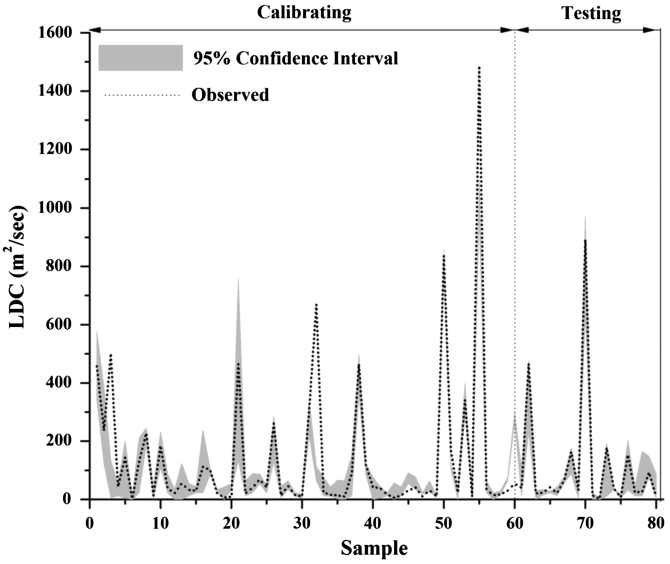

Fig. 6. 95% confidence intervals for the estimated LDC during calibration and testing steps using the SVM model

the same data sets. Table 4 summarizes the results of some studies on predicting LDC using AI techniques. The table indicates that the results, reported in previous studies on LDC under similar conditions, are highly variable, implying high uncertainty. Therefore, it is important to estimate the uncertainty in ANN, ANFIS, and SVM techniques. It can also be seen clearly from Figs. 4–6 that the predicted LDC values contain high uncertainty as compared with the observed ones. The uncertainty is illustrated by the 95% confidence interval for ANN, ANFIS, and SVM models. It should be noted that the 95% confidence interval of the models must be zero (d-factor equal to 0) if the uncertainty in the models is zero. However, the results from earlier studies, Table 4, and Figs. 4–6, show that the ANN, ANFIS, and SVM models involve high uncertainty in predicting LDC in natural rivers. Figs. 4–6 generally indicate that the ANN and ANFIS models have a wider 95PPU band than does the SVM model. It means that these two models have higher uncertainties than the SVM model. More investigations indicate that the d-factor parameter for the ANN and ANFIS models is larger than that for the SVM model, implying that these two models do not perform well in the prediction of LDC as compared with the SVM model. It is also observable that the number of predictions bracketed by 95PPU in all three models is acceptable in both the calibrating and testing steps. Therefore, the performance of the SVM model is superior to the other two models. In other words, the SVM model is more reliable than the ANN and ANFIS models for LDC prediction.

Table 4. Reported Results of Some Studies for Predicting LDC Using AI Techniques

Number	Reference	Model	R^2 in testing step
1	Tayfur and Singh (2005)	ANN	0.69
2	Toprak and Savci (2007)	ANFIS	0.99
3	Toprak and Cigizoglu (2008)	RBF	0.93
4	Toprak and Cigizoglu (2008)	GRNN	0.84
5	Toprak and Cigizoglu (2008)	FFBP	0.98
6	Riahi-Madvar et al. (2009)	ANFIS	0.91
7	Noori et al. (2009)	SVM	0.53
8	Noori et al. (2009)	ANFIS	0.50
9	Adarsh (2010)	SVM	0.76
10	Azamathulla and Wu (2011)	SVM	0.81
11	Present research	ANN	0.82
12	Present research	ANFIS	0.83
13	Present research	SVM	0.90

Note: FFBP = feed-forward error back-propagation neural networks; GRNN = generalized regression neural networks (GRNN); RBF = radial basis function-based neural networks.

Summary and Conclusions

While almost the same data sets and AI techniques were used for the prediction of LDC, significantly different results are often reported in the literature. It means that the LDC values estimated using AI techniques are highly uncertain. Therefore, uncertainty analysis of AI techniques was presented in this study to answer questions like, "How reliable are the predicted results of AI techniques?" or "How high is the uncertainty of AI techniques in predicting the LDC in natural rivers?" More specifically, the paper demonstrated a pioneering effort in the uncertainty analysis of AI models.

Artificial intelligent modeling of LDC in natural rivers was carried out in this paper using 100 data sets involving geometrical and hydraulic characteristics measured in more than 30 streams in the United States. Forward selection and gamma test techniques were used to sort the model input variables based on their importance to the LDC prediction with AI models. Thereafter, three best-trained models, based on ANN, ANFIS, and SVM techniques, were selected for predicting LDC in natural streams. Finally, an uncertainty analysis was carried out for the three models, individually. Results showed that the gamma test (nonlinear technique) was better than the forward selection (linear technique) for selecting input variables based on their importance. The SVM model was found to be superior in predicting LDC due to low uncertainty as compared with those in the ANN and ANFIS models, while the ANFIS model performs better than the ANN model. The LDC values, predicted with the SVM model, are acceptable due to the following reasons. First, the existence of fewer model input parameters to be optimized in the SVM model increases the actual performance of the model. Second, the SVM model has a strong theoretical background, making it more certain than the ANN and ANFIS models. Third, the SVM and ANFIS models can perform better than the ANN model, when fed with fewer data. Fourth, regarding the superiority of ANFIS to the ANN model, it should be noted that contrary to ANN, the ANFIS model is based on a fuzzy logic which causes it to face less uncertainty than the ANN model.

References

Abbaspour, K. C., et al. (2007). "Modeling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT." *J. Hydrol.*, 333(2–4), 413–430.

Abe, S. (2005). *Support vector machines for pattern classification*, Springer, London.

Adarsh, S. (2010). "Prediction of longitudinal dispersion coefficient in natural channels using soft computing techniques." *Sci. Iran.*, 17(5), 363–371.

Agalbjorn, S., Koncar, N., and Jones, A. J. (1997). "A note on the gamma test." *Neural Comput. Appl.*, 5(3), 131–133.

Aqil, M., Kita, I., Yano, A., and Nishiyama, S. (2007). "Analysis and prediction of flow from local source in a river basin using a neuro-fuzzy modeling tool." *J. Environ. Manage.*, 85(1), 215–223.

Azamathulla, H. M., and Ghani, A. A. (2011). "Genetic programming for predicting longitudinal dispersion coefficients in streams." *Water Resour. Manage.*, 25(6), 1537–1544.

Azamathulla, H. M., and Wu, F. C. (2011). "Support vector machine approach for longitudinal dispersion coefficients in natural streams." *Appl. Soft Comput.*, 11(2), 2902–2905.
Bencala, K. E., and Walters, R. A. (1983). "Simulation of solute transport in a mountain pool-and-riffle stream: A transient storage model." *Water Resour. Res.*, 19(3), 718–724.
Brown, M., and Harris, C. (1994). *Neuro-fuzzy adaptive modeling and control*, Prentice-Hall, NJ.
Chapra, S. C. (1997). *Surface water quality modeling*, McGraw-Hill, Boston.
Chen, S., Hong, X., Harris, C. J., and Sharkey, P. M. (2004). "Sparse modeling using orthogonal forward regression with PRESS statistic and regularization." *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, 34(2), 898–911.
Clark, J. F., Schlosser, P., Stute, M., and Simpson, H. J. (1996). "SF6-3He tracer release experiment: A new method of determining longitudinal dispersion coefficients in large rivers." *Environ. Sci. Technol.*, 30(5), 1527–1532.
Clark, M. P., and Slater, A. G. (2006). "Probabilistic quantitative precipitation estimation in complex terrain." *J. Hydrometeorol.*, 7(1), 3–22.
Coulibaly, P., Ancti, F., and Bobee, B. (2000). "Daily reservoir inflow forecasting using artificial neural networks with stopped training approach." *J. Hydrol.*, 230(3–4), 244–257.
Deng, Z., Bengtsson, L., Singh, V. P., and Adrian, D. D. (2002). "Longitudinal dispersion coefficient in single-channel streams." *J. Hydraul. Eng.*, 10.1061/(ASCE)0733-9429(2002)128:10(901), 901–916.
Deng, Z., and Jung, H.-S. (2009). "Variable residence time-based model for solute transport in streams." *Water Resour. Res.*, 45(3), W03415.
Deng, Z., Jung, H.-S., and Ghimire, B. (2010). "Effect of channel size on solute residence time distributions in rivers." *Adv. Water Resour.*, 33(9), 1118–1127.
Di Baldassarre, G., and Montanari, A. (2009). "Uncertainty in river discharge observations: A quantitative analysis." *Hydrol. Earth Syst. Sci.*, 13(6), 913–921.
Durrant, P. J. (2001). "winGamma™: A non-linear data analysis and modeling tool with applications to flood prediction." Ph.D. thesis, Dept. of Computer Science, Cardiff Univ., Wales, U.K.
Elder, J. W. (1959). "The dispersion of a marked fluid in turbulent shear flow." *J. Fluid Mech.*, 5(04), 544–560.
Etemad-Shahidi, A., and Taghipour, M. (2012). "Predicting longitudinal dispersion coefficient in natural streams using M5′ model tree." *J. Hydraul. Eng.*, 10.1061/(ASCE)HY.1943-7900.0000550, 542–554.
Evans, D., and Jones, A. J. (2002). "A proof of the gamma test." *Proc. Roy. Soc., Ser. A*, 458(2027), 2759–2799.
Fischer, B. H. (1975). "Discussion of 'Simple method for predicting dispersion in streams,' by R. S. McQuivey and T. N. Keefer." *J. Environ. Eng. Div.*, 101, 453–455.
Fischer, H. B., List, E. J., Koh, R. C. Y., Imberger, J., and Brooks, N. H. (1979). *Mixing in inland and coastal waters*, Academic, New York.
Graf, J. B. (1995). "Measured and predicted velocity and longitudinal dispersion at steady and unsteady flow, Colorado River, Glen canyon dam to Lake Mead." *J. Am. Water Resour. Assoc.*, 31(2), 265–281.
Haykin, S. (1994). *Neural networks: A comprehensive foundation*, Macmillan, New York.
Ho, D. T., Schlosser, P., and Caplow, T. (2002). "Determination of longitudinal dispersion coefficient and net advection in the tidal Hudson River with a large-scale, high-resolution SF6 tracer release experiment." *Environ. Sci. Technol.*, 36(15), 3234–3241.
Ho, D. T., Schlosser, P., Houghton, R. W., and Caplow, T. (2006). "Comparison of SF6 and fluoresce in as tracers for measuring transport processes in a large tidal river." *J. Environ. Eng.*, 10.1061/(ASCE)0733-9372(2006)132:12(1664), 1664–1669.
Hornik, K. (1989). "Multilayer feed forward networks are universal approximators." *Neural Networks*, 2(5), 359–366.
Huber, P. J. (1981). *Robust statistics*, Wiley, New York.
Jang, J. S. R. (1993). "ANFIS: Adaptive-network-based fuzzy inference system." *IEEE Trans. Syst. Man Cybern.*, 23(3), 665–685.
Jang, J. S. R., Sun, C. T., and Mizutani, E. (1997). *Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence*, Prentice-Hall, Englewood Cliffs, NJ.
Kashefipour, M. S., and Falconer, R. A. (2002). "Longitudinal dispersion coefficients in natural channels." *Water Res.*, 36(6), 1596–1608.
Koussis, A. D., and Rodriguez-Mirasol, J. (1998). "Hydraulic estimation of dispersion coefficient for streams." *J. Hydraul. Eng.*, 10.1061/(ASCE)0733-9429(1998)124:3(317), 317–320.
Krzysztofowicz, R. (2002). "Bayesian system for probabilistic river stage forecasting." *J. Hydrol.*, 268(1–4), 16–40.
Liu, H. (1977). "Predicting dispersion coefficient of streams." *J. Environ. Eng. Div.*, 103, 59–69.
McMillan, H., Jackson, B., Clark, M., Kavetski, D., and Woods, R. (2011). "Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models Mendeley." *J. Hydrol.*, 400(1–2), 83–94.
McQuivey, R. S., and Keefer, T. N. (1974). "Simple method for predicting dispersion in streams." *J. Environ. Eng.*, 100, 997–1011.
Montanari, A., and Grossi, G. (2008). "Estimating the uncertainty of hydrological forecasts: A statistical approach." *Water Resour. Res.*, 44(12).
Nayak, P. C., Sudheer, K. P., Rangan, D. M., and Ramasastri, K. S. (2004). "A neuro-fuzzy computing technique for modeling hydrological time series." *J. Hydrol.*, 291(1–2), 52–66.
Noori, R., et al. (2011a). "Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction." *J. Hydrol.*, 401(3), 177–189.
Noori, R., Karbassi, A., Farokhnia, A., and Dehghani, M. (2009). "Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques." *Environ. Eng. Sci.*, 26(10), 1503–1510.
Noori, R., Karbassi, A. R., Mehdizadeh, H., and Sabahi, M. S. (2011b). "A framework development for predicting the longitudinal dispersion coefficient in natural streams using artificial neural network." *Environ. Prog. Sustainable*, 30(3), 439–449.
Rajeev, R. S., and Dutta, S. (2009). "Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm." *Hydrol. Res.*, 40(6), 544–552.
Riahi-Madvar, H., Ayyoubzadeh, S. A., Khadangi, E., and Ebadzadeh, M. M. (2009). "An expert system for predicting longitudinal dispersion coefficient in natural streams by using ANFIS." *Expert Syst. Appl.*, 36(4), 8589–8596.
Rutherford, J. C. (1994). *River mixing*, Wiley, New York.
Seo, I. W., and Bake, K. O. (2004). "Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams." *J. Hydraul. Eng.*, 10.1061/(ASCE)0733-9429(2004)130:3(227), 227–236.
Seo, I. W., and Cheong, T. S. (1998). "Predicting longitudinal dispersion coefficient in natural streams." *J. Hydraul. Eng.*, 10.1061/(ASCE)0733-9429(1998)124:1(25), 25–32.
Smart, P. L., and Laidlaw, I. M. S. (1977). "An evaluation of some fluorescent dyes for water tracing." *Water Resour. Res.*, 13(1), 15–33.
Tayfur, G. (2006). "Fuzzy, ANN, and regression models to predict longitudinal dispersion coefficient in natural streams." *Nordic Hydrol.*, 1, 1–23.
Tayfur, G., and Singh, V. P. (2005). "Predicting longitudinal dispersion coefficient in natural streams by artificial neural network." *J. Hydraul. Eng.*, 10.1061/(ASCE)0733-9429(2005)131:11(991), 991–1000.
Taylor, G. I. (1954). "The dispersion of matter in turbulent flow through a pipe." *Proc. R. Soc. London, Ser. A*, 223(1155), 446–468.
Toprak, Z. F., and Cigizoglu, H. K. (2008). "Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods." *Hydrol. Processes*, 22(20), 4106–4129.
Toprak, Z. F., and Savci, E. M. (2007). "Longitudinal dispersion coefficient modelling in natural channels using fuzzy logic." *Clean-Soil, Air, Water*, 35(6), 626–637.
Vapnik, V. N. (1998). *Statistical learning theory*, Wiley, New York.