AN ARTIFICIAL NEURAL NETWORK MODEL FOR PREDICTING LONGITUDINAL DISPERSION COEFFICIENTS IN RIVERS

Ali A.M. Gad

Lecturer, Civil Engineering Department, Faculty of Engineering,
Assiut University, Assiut, Egypt
Email: aligad10@yahoo.com

Abstract

This study presents an artificial neural network (ANN) model to predict the values of the longitudinal dispersion coefficient (D_l) in rivers from their main hydraulic parameters. The model can be considered as a useful aid to water quality monitoring in rivers. The ANN model is a relatively new promising technique which can make use of the river width, depth, velocity, and shear velocity for predicting D_l. The used ANN model is based on a back propagation algorithm to train a multi-layer feed-forward network. The proposed model was verified using 116 sets of field data collected from 62 streams ranging from straight manmade canals to sinuous natural rivers. The ANN model predicts D_l, where more than 83% of the calculated values range from 0.50 to 2.0 times the observed values in the field. A comparison of the ANN model estimates with the outputs of the most recent and accurate equations in the literature, for the longitudinal dispersion coefficient, using three different statistical methods for analysis, has shown that the accuracy of the ANN model compared favourably with other equations. Finally, a new accurate predictor for the values of D_l in polluted streams that based on readily measurable hydraulic quantities is presented.

Keywords*: Water quality; Dispersion coefficient; Rivers; Neural network modelling*

1. INTRODUCTION

The longitudinal dispersion of pollutants in rivers is important to practicing environmental engineers for designing outfalls or water intakes and for evaluating risks from accidental releases of hazardous contaminants. The ability of rivers or other surface water bodies to disperse added substances is measured by the dispersion coefficients. The longitudinal dispersion coefficient can be introduced as a measurement of the one-dimensional dispersion process described by the classical convection-dispersion equation as:

$$\frac{\partial C}{\partial t} = D_l \frac{\partial^2 C}{\partial x^2} - U \frac{\partial C}{\partial x} \qquad (1)$$

in which C is the cross-sectional averaged concentration; D_l the longitudinal dispersion coefficient, U the mean longitudinal velocity; t the time; and x the longitudinal coordinate oriented in the direction of the mean flow. As the transport process and hence the fate of pollutants in surface water bodies depend to a large extent on D_l, many theoretical and empirical formulations have been proposed to determine it. Since many of these studies have been dependent upon different assumptions in the flow conditions, the values obtained for D_l have often varied widely.

In open channel flow, Elder [1] presented the first-published analysis of D_l based on laboratory measurements and Taylor's [2] method, by assuming a logarithmic vertical-velocity distribution to give the well known equation:

$$D_l = \left(\frac{0.4041}{k^3} + \frac{k}{6}\right) HU_* \qquad \text{or} \qquad D_l = 5.93HU_* \tag{2}$$

where k is the von-Karman constant, which is approximately equal to 0.41; H the depth of flow; and U_* the bed shear-stress velocity which can be given as:

$$U_* = \sqrt{gRS} \tag{3}$$

in which g is the gravitational acceleration; R the hydraulic radius; and S the slope of the energy gradient line. However, it has been found that Elder's equation significantly underestimates the dispersion coefficient. Studies undertaken using many measured data sets for natural rivers have shown that the value of D_l / HU_* may vary from 8.6 to 7500, with values generally being much greater than Elder's equation constant 5.93, Fischer et al. [3]. Guymer and West [4] confirmed the importance of both vertical and transverse shear components of the longitudinal dispersion coefficient. Seo and Cheong [5] derived a new equation using dimensional and regression analysis, their equation can be written as:

$$\frac{D_l}{HU_*} = 5.915\left(\frac{W}{H}\right)^{0.62}\left(\frac{U}{U_*}\right)^{1.428} \tag{4}$$

in which W is the channel width. Recently, Kashefipour and Falconer [6] developed the most recent, common, and accurate equation available in the literature for predicting the longitudinal dispersion coefficient in rivers flows, using dimensional and regression analysis. Furthermore, their equation was linearly combined with Seo and cheong's equation, led to a further improved equation for predicting the longitudinal dispersion coefficient in riverine and channel flows, giving:

$$D_l = \left[7.428 + 1.775\left(\frac{W}{H}\right)^{0.62}\left(\frac{U_*}{U}\right)^{0.572}\right] HU\left(\frac{U}{U_*}\right) \tag{5}$$

The artificial neural network (ANN) method is an artificial intelligence technique that attempts to mimic the human brains way of solving problems. In recent years, artificial neural network (ANN) models have attracted researchers in many disciplines of science and engineering, since they are capable of correlating large and complex data sets without any prior knowledge of the relationships among them. Artificial neural networks are capable to learn and organize themselves by extracting patterns and concepts directly from historical data. The first recorded use of ANN modelling in the field of civil engineering occurred in the early of 1980, when the technique was applied to the optimisation of construction tasks, Flood and Kartam [7].

The overall objective of the present study is to device and evaluate an artificial neural network (ANN) model for predicting the longitudinal dispersion coefficient for rivers based on readily measurable hydraulic parameters, which is as accurate as or more accurate than the existing empirical equations.

2. THEORY

2.1 Factors Influencing The Dispersion Process in Rivers

Major factors which influence dispersion characteristics of pollutants in natural streams can be categorized into three groups, Seo and Cheong [5]: fluid properties, hydraulic characteristics of the streams, and geometric configurations. The fluid properties include fluid density (ρ), viscosity (μ), and so on. The cross-sectional mean velocity (U), bed shear-stress velocity (U_*), channel width (W), and depth of flow (H) can be included in the category of bulk hydraulic characteristics. The bed forms and sinuosity can be regarded as the geometric configurations. The dispersion coefficient can be related to these parameters as:

$$D_l = f(\rho,\ \mu,\ U,\ U_*,\ W,\ H,\ S_f,\ S_n) \tag{6}$$

in which S_f is the bed shape factor and S_n the sinuosity. Since the flow in natural rivers and channels is generally fully turbulent and rough, with Reynolds number effects generally being negligible, ρ and μ can be ignored as a first approximation. S_f and S_n are vertical and transverse irregularities in natural streams, respectively, they cause secondary currents and shear flow that affect the hydraulic mixing processes in streams. In this study, however, S_f and S_n were dropped because they represent parameters not easily collected for natural streams, and furthermore, the influences of them can be included in the friction term. Dimensional analysis shows that there are many different combinations of W, H, U, *and* U_*, which can lead to the same dimensions as D_l.

2.2 Overview of The Artificial Neural Network

The artificial neural network (ANN) modelling approach is a computer methodology that attempts to simulate some important features of the human nervous system; in

other words, the ability to solve problems by applying information gained from past experience to new problems. Analogous to human brain, an ANN model uses many simple computational elements, named artificial neurons, connected by variable weights. Although each neuron, alone, can only perform simple computations, the hierarchical organization of a network of interconnected neurons makes an ANN capable of performing complex tasks such as pattern classification and prediction.

Artificial neural network (ANN) models are generally grouped into two broad categories, feed-forward networks and feed-backward networks, according to the pattern of flow of the model input information within the architecture. In feed-forward networks, the neurons on the first layer send their output to the neurons on the second layer, but they do not receive any input back from the neurons on the second layer. The network prediction error information may, however, be propagated in a backward direction through the network. In feed-backward networks, recurrent loops exist within the architecture that permits the network to retain a short-term memory with respect to previous input information. Such information is incorporated into the current information processing, making feed-backward networks particularly useful for time-series modelling.

2.3 Description of The Artificial Neural Network Modelling

The artificial neural networks (ANNs) are a relatively new technique, which can be used to predict the longitudinal dispersion coefficient by building a multi-layer feed-forward network. As shown in Fig. 1, the network consists of an input layer consisting of neuron(s) of so called node(s) representing various input variable(s), the hidden layer(s) that consisting of many hidden neurons for each layer, and an output layer consisting of neuron(s) representing various output variable(s). The number of hidden layers and neurons on each hidden layer is determined by a trial and error process. The input neurons pass on the input signal values to the neurons on the first hidden layer unprocessed. The values are distributed to all neurons on the first hidden layer depending on the connection weights (w_{ij}) between the input neuron (i) and the hidden neuron (j). On the first hidden layer, each unit j receives incoming signals from every unit i on the input layer. Associated with each incoming signal (x_i) is a weight (w_{ji}). The effective incoming signal (s_j) to the unit j is the weighted sum of all the incoming signals as:

$$s_j = \sum_{i=1}^{i=n} w_{ji} x_i \tag{7}$$

in which n is the number of neurons on the input layer. The effective incoming signal, s_j, is passed through a non-linear activation function, called a transfer function, to produce the outgoing signal $(y_j = f(s_j))$ of the unit j. The most commonly used transfer function in a multi-layer perceptron network (MLP) is the sigmoid function. The characteristics of the sigmoid function are that it is bounded above and below, continuous and differentiable every where. In this study, the type of the hyperbolic

sigmoid function used for the ANN model in the hidden layers is the tansig function, which can be written as:

$$f(s_j) = \frac{2}{1 + \exp^{-2s_j}} - 1 \tag{8}$$

in which s_j can vary on the range $\pm\infty$, but $f(s_j)$ is bounded between -1 and 1. In an analogous manner the processed signals from the neurons on the first hidden layer are distributed to the neurons on the second hidden layer and so on till the last hidden later. In the output layer all weighted incoming signals from neurons on the last hidden layer are summed and processed using a linear transfer function of the purelin type, which can be written as:

$$f(S'_j) = S'_j \tag{9}$$

Finally, the actual and network outputs are compared and the square of error is computed and summed for all variable patterns. The computed error is propagated backward from the output neurons to the hidden neurons to the input neurons based on the gradient delta rule. The connection weights are then updated to minimize the total network error. Once the training process is satisfactory completed, the final weights are saved and used for the evaluation of the model in the testing phase.

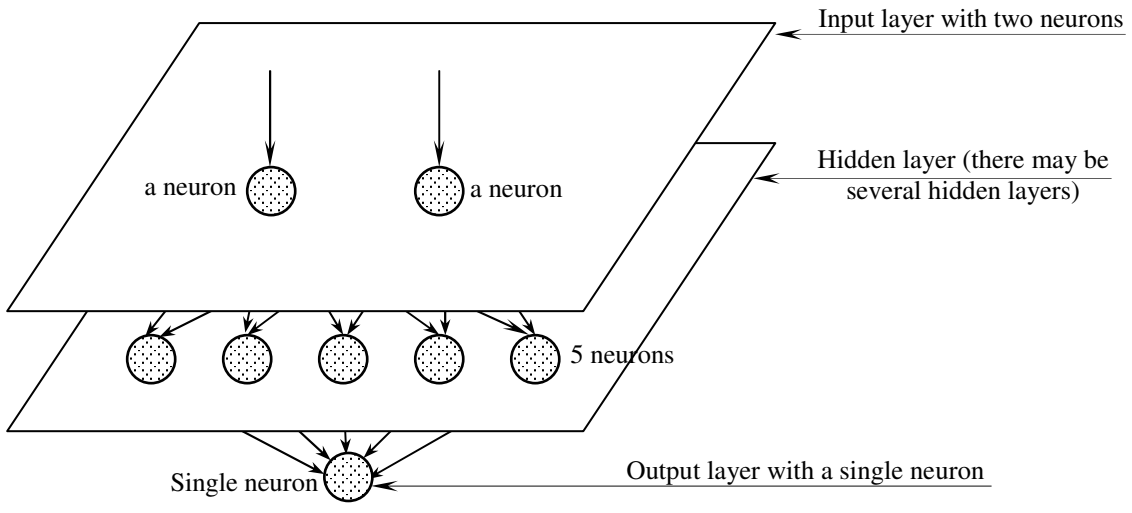

Figure 1. A schematic diagram of a simple feed-forward neural network

3. NUMERICAL PROCEDURE

3.1 Input and Output Parameters

In order to build up an ANN model to predict the longitudinal dispersion coefficient (D_l) for rivers, 116 data sets for more than 62 rivers and experimental flumes were collected from Fischer [3, 8, 9], Seo and Cheong [5], and McQuivey and Keefer [10]. The selection of the input variables had to be done in a way that enables the neural model to accomplish the task. The longitudinal dispersion coefficient has been related separately to all of the main hydraulic parameters. As shown in Fig. 2, D_l can be related to the channel width (Fig. 2a), depth of water (Fig. 2b), and velocity (Fig. 2c), whereas the data scattered in the $D_l - U_*$ plane (Fig. 2d). A perfect fit would result in a correlation coefficient (R^2) of a value of "1" while a value of "0" means very poor fit. The corresponding correlation coefficients, for the relationships between D_l and W, H, and U in Figs. 2a-c are 0.4816, 0.3781, and 0.4129, respectively. The figures demonstrate that D_l as an output parameters appears to have some dependency on all these variables, as the input parameters, even though the data somewhat scattered.

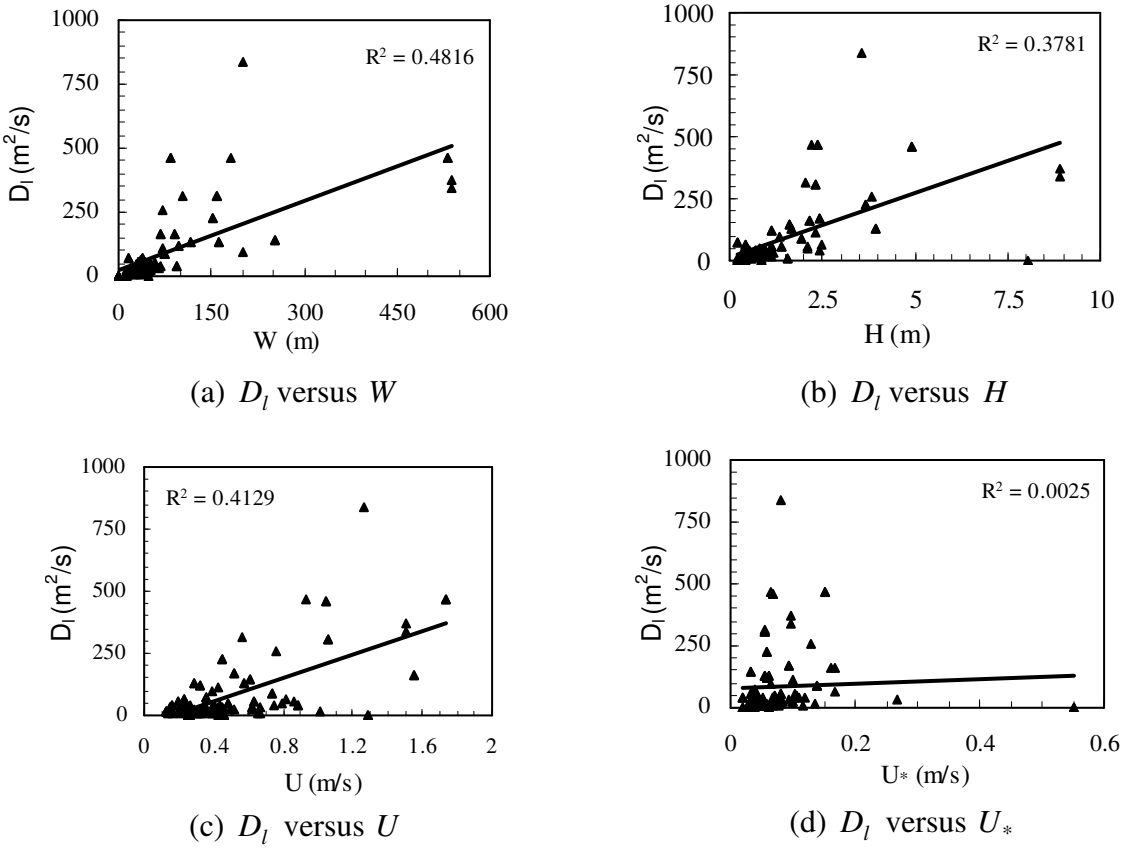

Figure 2. Relationship between D_l and (a)W, (b) H, (c) U, and (d) U_*

3.2 Model Training

In developing the ANN model for predicting D_l in polluted rivers based on their main hydraulic parameters, three configurations were evaluated: (i) training was carried out using the raw data values of the river width (W), depth (H), mean velocity (U) and shear velocity (U_*) as input parameters and the longitudinal dispersion coefficient (D_l) as an output parameter; (ii) the input parameters were putted as dimensionless values of W/H and U/U_* while D_l/HU_* as an output dimensionless parameter; and (iii) the dimensionless parameters of the second configuration were transformed to a logarithmic scale. Configuration (iii) yielded an optimal ANN model. The ANN configuration employed an input layer having 2 neurons, with one corresponding to each of $\log(W/H)$ and $\log(U/U_*)$ and an output layer consisting of a single neuron representing the output parameter $\log(D_l/HU_*)$.

In the training of the ANN model 69 data sets for 37 natural rivers and experimental flumes are randomly picked among the available 116 data sets, i.e. 60% of the available data. In order to find the optimal network, several configurations were tried in which the number of hidden layers varied from 1 to 4 and the number of neurons on each hidden layer was varied from 5 to 30. Once a given neural network was trained using the input data sets, its performance was then evaluated using the same data sets. It was found that the optimal configuration for the network, 4 hidden layers with 30, 20, 10, and 5 neurons on each layer, respectively. All the computations are made with the MATLAB® software (Release 12.1) and its neural modelling application, Neural Network Toolbox (version 4.4).

Figure 3 illustrates a comparison between the observed values of D_l in the field and those predicted by the ANN model, Seo and Cheong model (Eq. 4), and Kashefipour and Falconer model (Eq. 5). As the equation of Seo and Cheong (Eq. 4) and the equation of Kashefipour and Falconer (Eq. 5) are the most recent, common, and frequently used in computing D_l, they were chosen for comparison. It is clear from the figure that the outputs of the ANN model agree with the field observations while the other equations roughly estimate D_l. It is a rational result because the ANN model was trained on both the input parameters, $\log(W/H)$ *and* $\log(U/U_*)$, at the same time with the output parameter, D_l, in the form of $\log(D_l/HU_*)$. Thus, another new data sets are required to verify and compare the ANN model with the other equations.

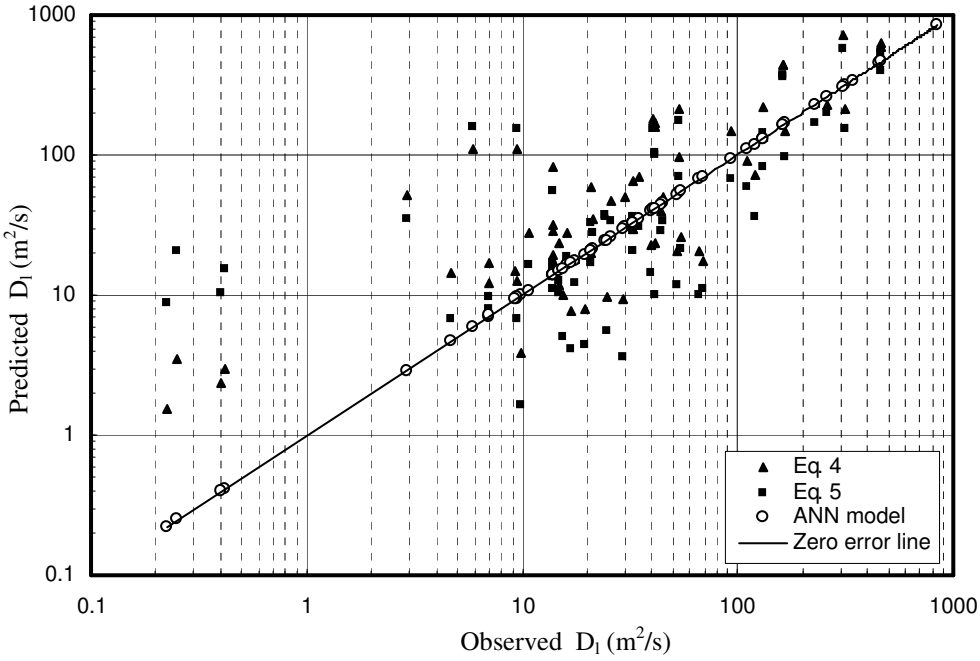

Figure 3. Comparison between the observed values of D_l and those predicted by Eq. (4), Eq. (5), and the ANN model through the training of the model

4. RESULTS AND DISCUSSION

In order to test the accuracy and feasibility of the ANN model for predicting the longitudinal dispersion coefficient (D_l) in natural streams, 47 data sets measured in 25 streams were used. The trained ANN model has never seen these data either in an input or output form and it was then used to predict the values of D_l as an output parameter based on the input parameters from the field measurements $(W,\ H,\ U,\ and\ U_*)$. The ANN model output values of D_l were then compared with both those values obtained from field observations and the outputs of the equations proposed by other investigators, as shown in Fig. 4. Compared to Eq. (4) and Eq. (5) outputs, it is clear from Fig. 4 that the developed ANN model improves the prediction of D_l and its estimates are more closer to the observed values from the field. In general, predicted longitudinal dispersion coefficients often deviate from observed ones by orders of magnitude. The deviation is attributed mainly to the inability to account for meandering and other non-uniform conditions of the rivers. Also, it is found from the statistical calculations that the ANN model predicts D_l with an accuracy in which 83% of the predicted values range from 0.50 to 2.0 times of the observed values form the same rivers and data sets. In comparison, values of 48.9% and 51.1% of the calculated values of D_l by Eq. (4) and Eq. (5), respectively, range from 0.50 to 2.0 times the observed ones.

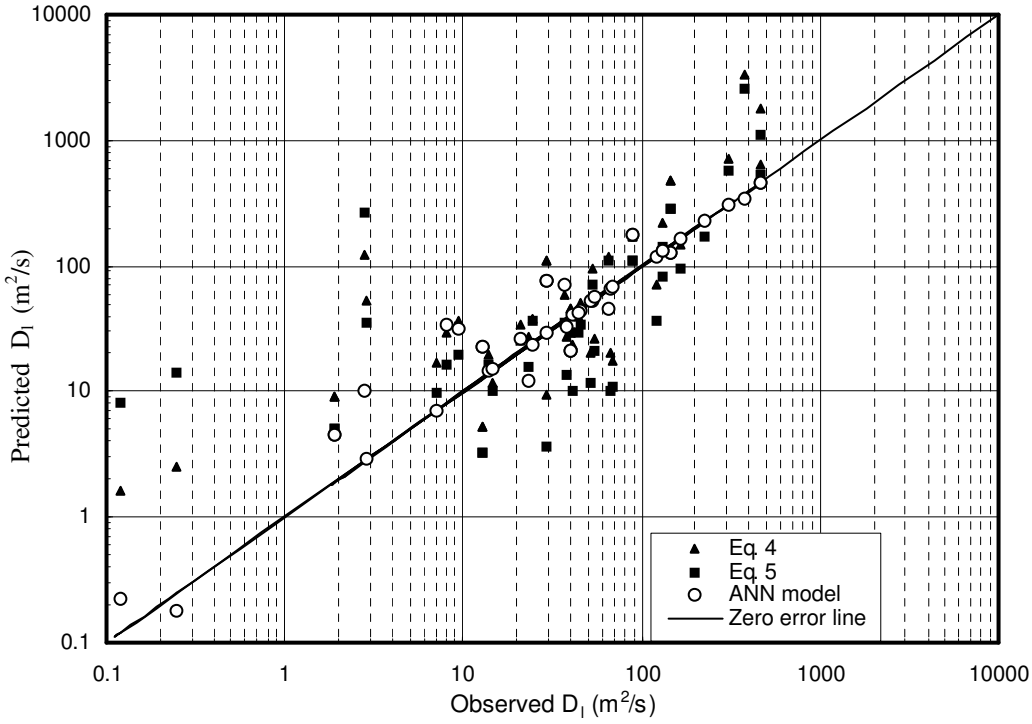

Figure 4. Comparison between the observed values of D_l and those predicted by Eq. (4), Eq. (5), and the ANN model through the testing of the model

Also, for statistical comparisons the coefficient of determination (C_d) and the mean relative absolute error $(MRAE)$ are used to compare the performance of the ANN model with both Eq. (4) and Eq. (5). The coefficient of determination is a statistical indicator that varies from $-\infty$ for bad models to 1.0 for good models. It represents the fraction of the total variance of the observed variable that is explained by the model. The coefficient of determination, C_d, is mathematically described as:

$$C_d = 1 - \frac{\sum_{i=1}^{i=n} (Dl_i^{Obs} - Dl_i^{\Pr d})^2}{\sum_{i=1}^{i=n} (Dl_i^{Obs} - \overline{D}l^{Obs})^2} \tag{10}$$

where the parameter $Dl_i^{\Pr d}$ represents the predicted output of D_l from the ANN model or any of the other two equations for a given input while Dl_i^{Obs} is the desired output for the same input which was observed in the field; $\overline{D}l^{Obs}$ the mean Dl_i^{Obs} values; and n the total number of data records. The another error measure that is used for comparison is the mean relative absolute error, $MRAE$, which can be defined as:

$$MRAE = \frac{1}{n} \sum_{i=1}^{i=n} \left| \frac{Dl_i^{Obs} - Dl_i^{\Pr d}}{Dl_i^{Obs}} \right| \tag{11}$$

Table (1) shows the values of the coefficient of determination, the mean relative absolute error, and the percent of the predicted D_l that lies in the range 0.50~2.0 times the observed field values, for the ANN model, Eq. (4), and Eq. (5). It is clear from the table that the ANN model performs much better than the other two equations for all values of the three error measures. It follows from the above investigations that the new optimised model of the ANN is capable of providing a superior prediction of the longitudinal dispersion coefficient for natural rivers.

Table 1. Comparison of the ANN model, Seo and Cheong model (Eq. 4), and Kashefipour and Falconer model (Eq. 5) using three statistical methods

Model Type	Coefficient of determination	Mean relative absolute error	Percent of $Dl_{Obs} / Dl_{\Pr d}$ values that ranges from 0.5 to 2.0
ANN model	0.9483	0.620	83.0 %
Eq. (4)	-17.377	2.837	48.9 %
Eq. (5)	-8.203	5.556	51.1 %

5. CONCLUSIONS

Using the promising technique of the artificial neural network (ANN), an accurate predictor for the longitudinal dispersion coefficient, D_l, for streams that is based on readily measurable hydraulic quantities is presented. The ANN model relates D_l to the main hydraulic parameters of the river width, depth, velocity, and bed shear-stress velocity. In developing the ANN model, 116 sets of field data collected from 62 streams ranging from straight manmade canals to sinuous natural rivers were used. The ANN estimates were compared with the outputs of two other existing equations frequently used to predict D_l in riverine flows, with the comparisons based on three different statistical methods. It was found that the new method of ANN has the least error and improves the prediction of D_l for natural rivers. The ANN model predicts D_l with an accuracy in which 83% of the calculated values range from 0.5 to 2.0 times the observed values in the field. The developed ANN model can be considered as a useful aid to water quality monitoring in rivers, which is as accurate as or more accurate than the most recent empirical equations.

REFERENCES

1. Elder, J.W., 1959, “The dispersion of a marked fluid in turbulent shear flow”, J. Fluid Mech., Cambridge, U. K., 5(4), 544-560.
2. Taylor, G.I., 1954, “The dispersion of matter in turbulent flow through a pipe”, Proc. Royal Soc., London, U.K., A223, 446-468.

3. Fischer, H.B., List, E.J., Koh, R.C.Y., Imberger, J., and Brooks, N.H., 1979, "Mixing in inland and coastal waters", New York: Academic Press, 483 pp.
4. Guymer, I., and West, J.R., 1992, "Longitudinal dispersion coefficients in estuary", J. Hydraul Eng., ASCE, 118, 718-734.
5. Seo, I.W., and Cheong, T.S., 1998, "Predicting longitudinal dispersion coefficient in natural streams", J. Hydraul. Eng., ASCE, 124, 25-32.
6. Kashefipour, S.M., and Falconer, R.A., 2002, "Longitudinal dispersion coefficients in natural channels", Water Research, 36, 1596-1608.
7. Flood, I., and Kartam, N., 1997, "Neural networks in civil engineering I: principles and understanding", J. computing in civil Eng., 8(2), 131-163.
8. Fischer, B.H., 1967, "The mechanics of dispersion in natural streams", Proceedings of the ASCE, J of Hydraul Div., 93, 187-216.
9. Fischer, B.H., (1968), "Dispersion predictions in natural streams", J. Sanit Eng. Div., ASCE, 94(5), 927-943.
10. McQuivey, R.S., and Keefer, T.N., 1974, "Simple method for predicting dispersion in streams", J. Environ Eng., ASCE, 100, 997-1011.