Environ Fluid Mech (2011) 11:247–261
DOI 10.1007/s10652-010-9175-y

ORIGINAL ARTICLE

Prediction of longitudinal dispersion coefficients in natural rivers using artificial neural network

Rajeev Ranjan Sahay

Received: 23 November 2009 / Accepted: 12 May 2010 / Published online: 30 May 2010

Abstract An artificial neural network (ANN) model is developed for predicting the longitudinal dispersion coefficient in natural rivers. The model uses few rivers' hydraulic and geometric characteristics, that are readily available, as the model input, and the target output is the longitudinal dispersion coefficient (K). For performance evaluation of the model, using published field data, predictions by the developed ANN model are compared with those of other reported important models. Based on various performance indices, it is concluded that the new model predicts the longitudinal dispersion coefficient more accurately. Sensitive analysis performed on input parameters indicates stream width, flow depth, stream sinuosity, flow velocity, and shear velocity to be the most influencing parameters for accurate prediction of the longitudinal dispersion coefficient.

Keywords Dispersion · Longitudinal dispersion coefficient · Artificial neural network · Streams · Pollutant dispersion · River flow

1 Introduction

The mixing process of pollutants in natural rivers or streams is complicated due to irregularities of velocity, bed configuration, development of secondary flow, dead-zone, and so on [9]. Pollutants undergo stages of mixing and dispersion as the flowing water transports them downstream. Advective and diffusive processes disperse them transversely and vertically, and once the cross-sectional and vertical mixing is complete, the process of longitudinal dispersion becomes important, intensity of which is measured by the longitudinal dispersion coefficient (K). The knowledge of the accurate value of K for a stream is essential for designing water intakes, outfalls, and treatment plants; and for assessing environmental impact due to accidental injection of a pollutant into the stream. Direct measurement of longitudinal dispersion coefficients, with the aid of concentration samples taken in upstream

R. R. Sahay (✉)
Department of Civil Engineering, Birla Institute of Technology, Mesra, Ranchi 835215, India
e-mail: rajeev_sahay@yahoo.com

and downstream cross-sections of a river, is rather seldom; K is usually estimated with the aid of theoretical and/or empirical expressions [20].

2 Previous studies

While deriving the following evolution equation for 1-D dispersion of tracer in a laminar pipe flow, Taylor first introduced the concept of K [25,26]:

$$A\frac{\partial C}{\partial t} = -UA\frac{\partial C}{\partial x} + \frac{\partial}{\partial x}\left(KA\frac{\partial C}{\partial x}\right) \tag{1}$$

where K is the longitudinal dispersion coefficient at distance x from the point of injection of the pollutant; A is the mean cross sectional flow area; C is the cross-sectional average concentration of the pollutant; U is the mean longitudinal flow velocity; and t is the time of observation.

Elder [3] extended Taylor's method for an open channel of infinite width in which he assumed logarithmic velocity profile in vertical direction and obtained a simple expression for K as

$$K = 5.93HU_* \tag{2}$$

where U_* is the bottom average shear friction velocity along the wetted perimeter of the cross-section, and H is the mean depth of flow.

Elder's expression was widely used for its simplicity and strong theoretical background; however, there were large discrepancies between the measured and the predicted values. Fischer [4] postulated that in natural streams, the transverse profile of the velocity is more important than the vertical profile for dispersion phenomena, and obtained an integral relation for the dispersion coefficient in natural streams having large width to depth ratios as

$$K = -\frac{1}{A}\int_0^W hu' \int_0^y \frac{1}{\varepsilon_t h} \int_0^y hu' \mathrm{d}y\mathrm{d}y\mathrm{d}y \tag{3}$$

where W is stream width at the free surface; h is the local depth of flow; u' is the difference between the local depth-averaged and the cross-sectional mean velocities; ε_t is the transverse, turbulent diffusion coefficient; and y is the lateral co-ordinate of the observation point from left bank of the stream.

The estimate of K through Eq. 3 is found to be inadequate, reasons for the discrepancy being the fact that no natural channel completely fulfills the assumptions inherent in the development of Eq. 3, as natural channels have bends, changes in shape, pools, and many other irregularities, all of which contribute significantly to the dispersion process. Equation 3 also requires elaborate transverse profiles of the velocity and the cross-sectional geometry that are not readily available to the field engineers. So Fischer et al. [5] developed a simpler expression for K by introducing a reasonable approximation for the triple integration, velocity deviation, and transverse turbulent diffusion coefficient

$$K = 0.011\frac{U^2W^2}{HU_*} \tag{4}$$

Using the similarity between the 1-D solute dispersion expression and the 1D flow expression, for the case when Froude number is less than 0.5, McQuivey and Keefer [14] developed another expression for K:

$$K = 0.058 \frac{HU}{S} \tag{5}$$

where S is downstream slope of the energy grade line.

Liu [12] derived a dispersion coefficient expression, taking into account the role of lateral velocity gradients in dispersion, in natural streams as

$$K = \alpha \frac{U^2 W^2}{HU_*} \tag{6}$$

where α is dependent upon the channel cross section, shape, and the velocity distribution across the stream, and can be estimated as (Godfrey and Frederick [6])

$$\alpha = 0.18 \left(\frac{U_*}{U}\right)^{1.5} \tag{7}$$

After analyzing data from their own experiments and from Nordin and Sabol [15], Iwasa and Aya [8] derived an empirical equation for K as

$$K = 2.0 HU_* \left(\frac{W}{H}\right)^{1.5} \tag{8}$$

Seo and Cheong [21], using the one-step Huber method for nonlinear multi-regression on published data, gave the following empirical expression for prediction of the longitudinal dispersion coefficient K and showed its superiority over other reported expressions

$$K = 5.915\,(HU_*) \left(\frac{W}{H}\right)^{0.62} \left(\frac{U}{U_*}\right)^{1.428} \tag{9}$$

Deng et al. [2] developed an analytical method to determine K for stable and straight alluvial rivers. They first derived a new transverse profile equation (ε_t) for channel shape and flow depth, and determined the lateral distribution of the deviation of the local depth mean velocity from cross-sectional mean velocity (u'), and substituted their values in Eq. 3 to predict K as

$$K = 0.15 \left(\frac{HU_*}{8\varepsilon_t}\right) \left(\frac{W}{H}\right)^{5/3} \left(\frac{U}{U_*}\right)^{2} \tag{10}$$

where

$$\varepsilon_t = 0.145 + \left(\frac{1}{3520.0}\right) \left(\frac{W}{H}\right)^{1.38} \left(\frac{U}{U_*}\right) \tag{11}$$

Seo and Baek [20] developed a new equation, similar to Eq. 4, for K that is based on the beta function for the transverse velocity profile:

$$K = \gamma \frac{U^2 W^2}{HU_*} \tag{12}$$

where γ is a constant, determination of which is rather difficult.

Using dimensional and regression analyses on 81 sets of data obtained from 30 rivers in USA, Kashefipour and Falconer [10] developed another expression to predict K in rivers:

$$K = 10.612\,(HU) \left(\frac{U}{U_*}\right) \tag{13}$$

Sahay and Dutta [19], using genetic algorithm on datasets of Deng et al. [2], derived expression for K as

$$\frac{K}{HU_*} = 2\left(\frac{W}{H}\right)^{0.96}\left(\frac{U}{U_*}\right)^{1.25} \tag{14}$$

Several other investigators have also attempted to provide estimates of K (Sukhodolov et al. [23], Koussis and Rodriguez-Mirazol [11], Piasecki and Katopodes [18], and Patil et al. [17]), to name a few.

The above discussion shows that most of the studies reported so far are based on specific assumptions and channel conditions, and therefore, for the same stream, the predicted longitudinal dispersion coefficient varies widely. Thus, a model, which has the general applicability and accuracy, is needed. In recent years, the artificial neural network (ANN) methods have been successfully applied to a number of multivariate forecasting problems in the field to hydrology. Important advantages of ANN models over other predictive models are their ability to model a phenomenon even if data is scanty, noisy, and filled with outliers. ASCE task committee [1] has given a summary of applications of ANNs in hydrology.

ANN application in the prediction of K is few. Tayfur and Singh [24], and Toprak and Cigizoglu [27] came up with studies in which they predicted the values of K using ANN. Their studies showed that ANNs were predicting K with higher accuracy. However, their studies had some deficiencies, as they did not consider all important input parameters affecting longitudinal dispersion, for; they left out a few relevant input parameters. Tayfur and Singh [24] developed five ANN models, and the input variables in their best performing model were mean depth of flow (H), stream width (W), and mean longitudinal flow velocity (U). They ignored average bed shear velocity (U_*) and channel sinuosity (σ). The present study shows that inclusion of these parameters (U_*, σ) improves the performance of the ANN model significantly. Toprak and Cigizoglu [27] employed three different architectures of ANN for prediction of K in which they used 65 published datasets. Their best ANN model was feed forward back propagation (FFBP) with W, H, U, and U_* as input parameters. They also ignored channel sinuosity (σ).

The present work aims at making investigation on the cases, which Tayfur and Singh [24], and Toprak and Cigizoglu [27] have overlooked.

3 ANN approach

For a few years, ANNs have become widely used forecasting tool. The working of an ANN mimics that of our brain. As the brain witnesses different causes with their effects and is experienced to forecast the result for an unforeseen but similar cause, an ANN, which is circuits of receiving and processing neurons, is trained for given pattern of inputs and outputs. The ability of ANNs to capture relationships from given patterns has enabled investigators to employ them in various hydraulic and hydrologic problems such as modeling river runoff, stream water level, river salinity, evapotranspiration, ground water table fluctuation, reservoir operation, to name a few.

Over the years, many variants of ANN are developed. However, three layer FFBP algorithm, because of the simplicity in its use and comprehension, is the most widely used ANN network. A typical FFBP has g, n, and m nodes or neurons in the input, hidden, and output layers, respectively. Neurons of a layer are connected to every neuron of the succeeding layer but are not connected among themselves. The parameters associated with each of the connections nodes, called weights, signify relative importance of the connections. All connections are feed forward, i.e., they allow information transfer only from an earlier layer to the next consecutive layers. Each node j receives incoming signals from every node i in the previous layer magnified by the weight of the connection (θ_{ji}). The effective incoming signal (a_j) to node j is the weighted sum of all the incoming signals plus a bias (b_j), i.e.;

$$a_j = b_j + \sum \theta_{ji} z_i \tag{15}$$

The effective incoming signal, a_j, is passed through a transfer function f (also called an activation function) to produce the outgoing signal (y_j) from the node j. In this study, sigmoid nonlinear function is used in the hidden and the output layers, i.e.,

$$y_j = f\left(a_j\right) = \frac{1}{1 + \exp\left(-a_j\right)} \tag{16}$$

The data variables are standardized before applying the ANN methodology. The method of standardization is dependent upon the transfer function used. In this work, sigmoid transfer function has been used in hidden and output layers, so data variables are standardized on to the range [0, 1]. Though there is no fixed rule, in this study the following standardization function has been used

$$z_i = \frac{x_i}{x_{\max}} \tag{17}$$

where z_i is the standardized data value, obtained by dividing the data x_i by the maximum of data values $x_{\max}$ for the neuron i.

The created network is first trained for the given data patterns. The experimental datasets include training and testing sets. The training, also called learning, determines the best interconnection weights and biases. It is accomplished by using known inputs and outputs in some ordered manner, adjusting the interconnection weights and biases until the desired outputs are achieved. There are many training algorithm. In the work, Levenberg–Marquardt optimization algorithm (Shepherd [22]) has been employed. The working principle of ANNs can be found in Haykin [7] or in any other book on the subject. The number of nodes in the hidden layer is determined by trial and error. ANNs with range of 2–12 hidden nodes are tried on the training datasets. The ANN model producing the best performance when applied to the testing datasets is selected. In this study, the ANN models are implemented using the software MATLAB 7 (Mathswork [13]). The mean square error (MSE) determines the performance index for training and testing of the model. The MSE is given by

$$\mathrm{MSE} = \frac{1}{N} \sum_{p=1}^{p=N} (K_{\mathrm{meas}} - K_{\mathrm{pred}})^2 \tag{18}$$

where K_{meas}, K_{pred} are measured and predicted longitudinal dispersion coefficients, respectively, for the pth pattern or dataset; and N is the total number of pattern.

For training and testing of ANNs of the present work, 71 field datasets, which were utilized by Tayfur and Singh [24] for their study, have been used. Same datasets have already been used by many other investigators, which facilitated good comparison of models. The geometric and hydraulic properties of the rivers along with measured longitudinal dispersion coefficients are given in Table 1. Tables 2 and 3 summarize statistical information on the measured datasets. Care has been taken while choosing datasets such that the range of mean values of training and testing sets are comparable. The channel shape parameter (β) is defined as logarithmic value of the ratio of stream width to flow depth; and the channel sinuosity (σ) of a stream is defined as the ratio of the channel length to the valley length. It's value is unity for straight rivers. Some values of σ from Table 1 are missing, they are

Table 1 Measured longitudinal dispersion coefficient in streams

S. no.	Stream	Width W (m)	Depth H (m)	Velocity U (m/s)	Sh. vel. U_* (m/s)	β	σ	W/H	U/U_*	Meas. Disp. K (m^2/s)
1	Amita River	37.0	0.81	0.29	0.070	3.82	–	45.68	4.14	23.2
2	Amita River	42.0	0.80	0.42	0.069	3.96	–	52.50	6.09	30.2
3	Antietam Creek, Md.	12.8	0.30	0.42	0.057	3.75	1.40	42.67	7.37	17.5
4	Antietam Creek, Md.	24.1	0.98	0.59	0.098	3.2	2.25	24.59	6.02	101.5
5	Antietam Creek, Md.*	11.9	0.66	0.43	0.085	2.89	2.25	18.03	5.06	20.9
6	Antietam Creek, Md.	21.0	0.48	0.62	0.069	3.78	1.26	43.75	8.99	25.9
7	Bayou Anacoco	20.0	0.42	0.29	0.045	3.66	1.41	47.62	6.44	13.9
8	Bayou Anacoco	17.5	0.45	0.32	0.024	3.32	1.41	38.89	13.33	5.8
9	Bayou Anacoco	25.9	0.94	0.34	0.067	3.69	1.41	27.55	5.07	32.5
10	Bayou Anacoco	36.6	0.91	0.40	0.067	3.86	1.41	40.22	5.97	39.5
11	Bayou Bartholomew, La	33.4	1.40	0.20	0.031	3.17	2.46	23.86	6.45	54.7
12	Bear Creek, Colo.*	13.7	0.85	1.29	0.553	2.78	1.08	16.12	2.33	2.9
13	Chattahoochee River, Ga*	75.6	1.95	0.74	0.138	3.66	1.27	38.77	5.36	88.9
14	Chattahoochee River, Ga	91.9	2.44	0.52	0.094	3.63	1.57	37.66	5.53	166.9
15	Clinch River, Va	48.5	1.16	0.21	0.069	3.73	1.25	41.81	3.04	14.8
16	Clinch River, Va*	28.7	0.61	0.35	0.069	3.85	1.14	47.05	5.07	10.7
17	Clinch River, Va	57.9	2.45	0.75	0.104	3.16	1.14	23.63	7.21	40.5
18	Clinch River, Va*	53.2	2.41	0.66	0.107	3.1	1.14	22.07	6.17	36.9
19	Comite River	13.0	0.26	0.31	0.044	3.91	1.31	50.00	7.05	7.0
20	Comite River	16.0	0.43	0.37	0.056	3.62	1.31	37.21	6.61	13.9
21	Comite River, La	15.7	0.23	0.36	0.039	4.22	1.31	68.26	9.23	69.0
22	Conococheague Creek, Md.	42.2	0.69	0.23	0.064	4.11	2.25	61.16	3.59	40.8
23	Conococheague Creek, Md.	49.7	0.41	0.15	0.081	4.8	2.25	121.22	1.85	29.3
24	Conococheague Creek, Md.*	43.0	1.13	0.63	0.081	3.64	1.31	38.05	7.78	53.3
25	Copper Creep, Va.	16.7	0.49	0.20	0.080	3.53	2.54	34.08	2.50	16.8
26	Copper Creep, Va.	18.3	0.38	0.15	0.116	3.88	2.54	48.16	1.29	20.7
27	Copper Creep, Va.	16.8	0.47	0.24	0.080	3.58	2.54	35.74	3.00	24.6
28	Copper Creep, Va.	19.6	0.84	0.49	0.101	3.15	1.26	23.33	4.85	20.8
29	Difficult Run, Va.	14.5	0.31	0.25	0.062	3.85	1.09	46.77	4.03	1.9
30	John Day River, Ore.*	25.0	0.58	1.01	0.140	3.76	1.08	43.10	7.21	13.9
31	John Day River, Ore.*	34.1	2.47	0.82	0.180	2.62	1.89	13.81	4.56	65.0
32	Little Pincy Creek, Md.	15.9	0.22	0.39	0.053	4.28	1.13	72.27	7.36	7.1
33	Minnesota River	80.0	2.74	0.03	0.002	3.37	–	29.20	14.17	22.3
34	Minnesota River	80.0	2.74	0.14	0.010	3.37	–	29.20	14.43	34.9
35	Mississippi River, La*	711.2	19.94	0.56	0.041	3.58	1.44	35.67	13.66	237.2
36	Mississippi River, Mo*	533.4	4.94	1.05	0.069	4.68	1.38	107.98	15.22	457.7
37	Mississippi River, Mo*	537.4	8.90	1.51	0.097	4.1	1.38	60.38	15.57	374.1
38	Missouri River	183.0	2.33	0.89	0.066	4.36	1.35	78.54	13.48	465.0
39	Missouri River*	197.0	3.11	1.53	0.078	4.15	1.35	63.34	19.62	892.0
40	Missouri River	201.0	3.56	1.28	0.084	4.03	1.35	56.46	15.24	837.0

Table 1 continued

S. no.	Stream	Width W (m)	Depth H (m)	Velocity U (m/s)	Sh. vel. U_* (m/s)	β	σ	W/H	U/U_*	Meas. Disp. K (m^2/s)
41	Monocacy River, Md.*	48.7	0.55	0.26	0.052	4.48	1.28	88.55	5.00	37.8
42	Monocacy River, Md.*	93.0	0.71	0.16	0.046	4.88	1.28	130.99	3.48	41.4
43	Monocacy River, Md.	51.2	0.65	0.62	0.044	4.37	1.28	78.77	14.09	29.6
44	Monocacy River, Md.	97.5	1.15	0.32	0.058	4.44	1.61	84.78	5.52	119.8
45	Monocacy River, Md.	40.5	0.41	0.23	0.040	4.59	1.61	98.78	5.75	66.5
46	Muddy River	13.0	0.81	0.37	0.081	2.77	–	16.05	4.57	13.9
47	Muddy River	20.0	1.20	0.45	0.099	2.82	–	16.67	4.55	32.5
48	Nooksack River	86.0	2.93	1.20	0.530	3.38	1.30	29.35	2.26	153.0
49	Nooksack River	64.0	0.76	0.67	0.268	4.43	1.30	84.21	2.50	34.8
50	Powell River, Tenn.*	36.8	0.87	0.13	0.054	3.74	2.20	42.30	2.41	15.5
51	Red River, La	253.6	1.62	0.61	0.032	5.05	1.20	156.54	19.06	143.8
52	Red River, La	161.5	3.96	0.29	0.060	3.93	1.44	40.78	4.83	130.5
53	Red River, La	152.4	3.66	0.45	0.057	3.73	1.44	41.64	7.89	227.6
54	Red River, La	155.1	1.74	0.47	0.036	4.49	1.24	89.14	13.06	177.7
55	Sabina River, La.	116.4	1.65	0.58	0.054	4.26	1.19	70.55	10.74	131.3
56	Sabina River, La.*	160.3	2.32	1.06	0.054	4.24	1.17	69.09	19.63	308.9
57	Sabina River, Tex.*	14.2	0.50	0.13	0.037	3.35	2.53	28.40	3.51	12.8
58	Sabina River, Tex.	12.2	0.51	0.23	0.030	3.17	2.05	23.92	7.67	14.7
59	Sabina River, Tex.	21.3	0.93	0.36	0.035	3.13	1.47	22.90	10.29	24.2
60	Salt Creek, Nebr.	32.0	0.50	0.24	0.038	4.16	1.38	64.00	6.32	52.2
61	Susquehanna River	203.0	1.35	0.39	0.065	5.01	1.13	150.37	6.00	92.9
62	Tangipahoa River, La	31.4	0.81	0.48	0.072	3.66	1.46	38.77	6.67	45.1
63	Tangipahoa River, La*	29.9	0.40	0.34	0.020	4.31	1.46	74.75	17.00	44.0
64	Tickfau River, La	15.0	0.59	0.27	0.080	3.23	1.75	25.42	3.38	10.3
65	White River*	67.0	0.59	0.35	0.044	4.73	–	113.56	7.95	30.2
66	Wind/Big. River, Wyo	44.2	1.37	0.99	0.142	3.48	1.56	32.26	6.97	184.6
67	Wind/Big. River, Wyo	85.3	2.38	1.74	0.153	3.58	1.56	35.84	11.37	464.6
68	Wind/Big. River, Wyo*	59.4	1.10	0.88	0.119	3.99	1.18	54.00	7.39	41.8
69	Wind/Big. River, Wyo	68.6	2.16	1.55	0.168	3.46	1.18	31.76	9.23	162.6
70	Yadkin River, N.C.	70.1	2.35	0.43	0.101	3.39	2.17	29.83	4.26	111.5
71	Yadkin River, N.C.	71.6	3.84	0.76	0.128	2.92	2.17	18.65	5.94	260.1

* Dataset used for testing

assumed to be 1. As can be seen from Tables 2 and 3, these datasets represent wide range of parameters. In order to check the suitability of the ANN model for predicting K, five different models are constructed. For training of ANNs, 51 out of the 71 datasets are utilized. Each ANN is trained with 0.02 learning rate, and up to 10000 iterations, or when MSE started rising. After completion of the training, the network is tested for prediction of K by using the remaining 20 datasets. The architecture of the best performing ANN model is shown in Fig. 1. It has five variables, W, H, U, U_*, and σ, in the input layer, eight neurons in the hidden layer, and one neuron in the output layer. The interconnection weights and biases for the model are given in the Appendix I.

Table 2 Average values for the whole, training and testing datasets for ANNs

Set	U_{mean} (m/s)	$U_{*\text{mean}}$ (m/s)	$(W/H)_{\text{mean}}$ (m)	β_{mean}	σ_{mean}	K_{mean} (m^2/s)	$\%W/H > 50$	$\%K > 100$
Whole	0.541	0.088	51.68	3.785	1.54	107.7	35.2	29.6
Training	0.481	0.082	50.26	3.769	1.57	95.3	31.4	31.4
Testing	0.690	0.100	55.30	3.830	1.46	139.3	45.0	25.0

Note: U_{mean} mean value of U, U_{*mean} mean value of U_*, $(W/H)_{mean}$ average value of (W/H), β_{mean} average value of β, σ_{mean} average value of σ, and K_{mean} average value of K

Table 3 Range values for the whole, training and testing datasets for ANNs

Set	$U_{\min}-U_{\max}$ (m/s)	$U_{*\min}-U_{*\max}$ (m/s)	$(WH)_{\min}-(W/H)_{\max}$	$\beta_{\min}-\beta_{\max}$	$\sigma_{\min}-\sigma_{\max}$	$K_{\min}-K_{\max}$ (m^2/s)
Whole	0.034−1.74	0.0024−0.55	13.81−156.54	2.62−5.05	1.54−2.54	1.9−892.0
Training	0.034−1.74	0.0024−0.53	16.05−156.54	2.77−5.05	1.09−2.54	1.9−837.0
Testing	0.130−1.53	0.0200−0.55	13.81−131.00	2.62−4.88	1.08−2.53	2.9−892.0

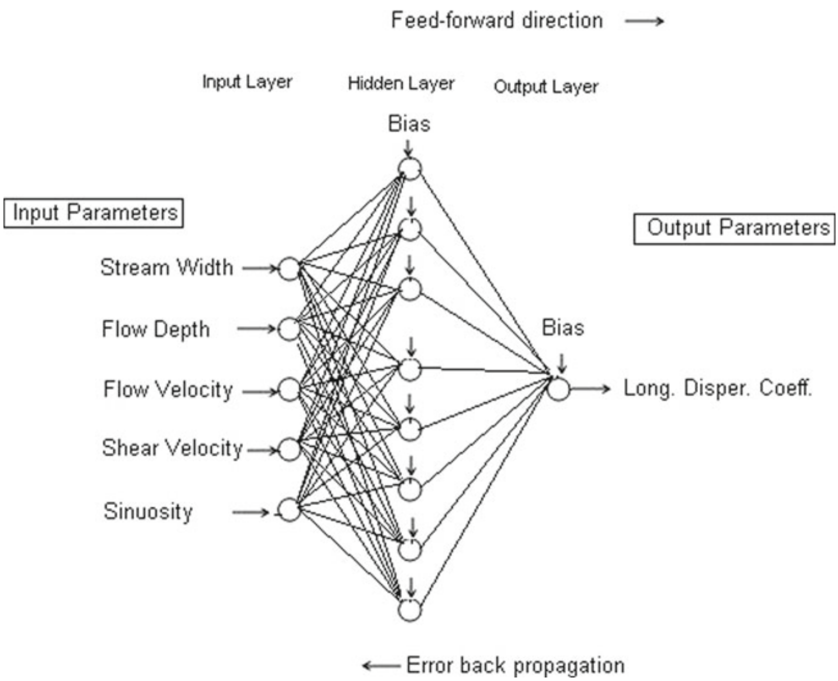

Fig. 1 The best ANN architecture of the present study for prediction of K

Figure 2 shows the predicted values of K for the training datasets by the best performing ANN model along with the measured K values. The model successfully predicted even extreme values of K; 837, 465, and 464.6 m^2/s as 821.3, 449.7, and 471.6 m^2/s, respectively. The solid line inclined at 45° shows over and under prediction. Although, the ANN model

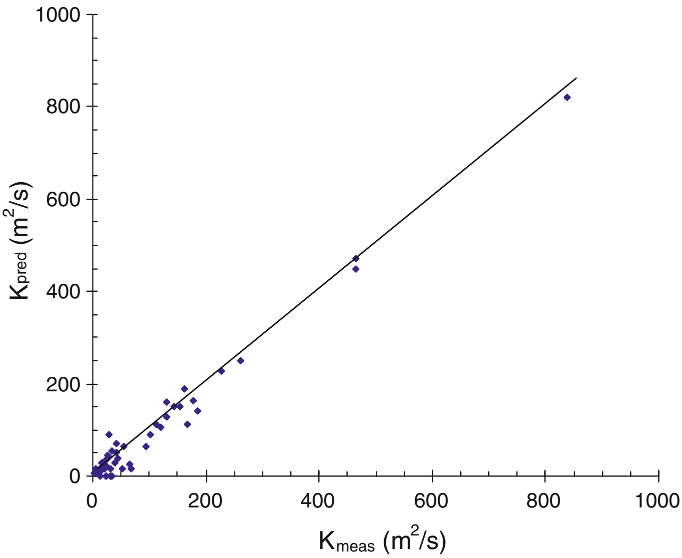

Fig. 2 Predicted and measured dispersion coefficient by the ANN model for the training dataset

prediction is evenly distributed around the inclined solid line, the number of underestimated values is a little more than the number of overestimated values. However, the overhaul coefficient of correlation (CC) between 51 measured and predicted K values at the end of the training period are 0.99, which indicates that the ANN model is satisfactorily trained.

4 Verification and comparison of the ANN model

For performance comparison, the predictions are made using 20 datasets employing different models. Models/expressions considered for comparison are the present ANN model (input datasets comprising W, H, U, U_*, and σ), Fischer (Eq. 4), Liu (Eq. 6), Iwasa and Aya (Eq. 8), Seo and Cheong (Eq. 9), Deng et al. (Eq. 10), Kashefipour and Falconer (Eq. 13), the ANN model of Tayfur and Singh [24], and Sahay and Dutta (Eq. 14). For brevity, the comparison models are denoted as Fischer, Liu, I–A, S–C, D–S–B, K–F, ANN (T–S), and S–D, respectively. Performance indices considered for comparing results from different models are the root mean square error (RMSE), the coefficient of correlation (CC), and the discrepancy ratio (DR). RMSE, CC, and DR are defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(K_{\mathrm{pred}} - K_{\mathrm{meas}}\right)^2}{N}} \tag{19}$$

$$\mathrm{CC} = \frac{\left[\sum_{1}^{N}\left(K_{\mathrm{meas}} - \overline{K}_{\mathrm{meas}}\right)\left(K_{\mathrm{pred}} - \overline{K}_{\mathrm{pred}}\right)\right]}{\sqrt{\sum\left(K_{\mathrm{meas}} - \overline{K}_{\mathrm{meas}}\right)^2 \sum\left(K_{\mathrm{pred}} - \overline{K}_{\mathrm{pred}}\right)^2}} \tag{20}$$

$$\mathrm{DR} = \log\frac{K_{\mathrm{pred}}}{K_{\mathrm{meas}}} \tag{21}$$

where $\overline{K}_{\text{meas}}$, $\overline{K}_{\text{pred}}$ are the mean measured and predicted longitudinal dispersion coefficients respectively; and N is the number of observations made.

From Eq. 21, DR $= 0$ implies exact matching between measured and predicted values; otherwise, there is either an over prediction (DR > 0, i.e., $K_{\text{pred}} > K_{\text{meas}}$) or under prediction (DR < 0, i.e., $K_{\text{pred}} < K_{\text{meas}}$). The performance of the present ANN model could not be compared with that of Toprak and Cigizoglu [27], as their test data and ours differed. However, Toprak and Cigizoglu [27] is best performing model, FFBP with H, W, U, and U_* as input, has RMSE equal to 115.2 m^2/s, which is larger than the present model's RMSE which is equal to 37.6 m^2/s. Moreover, Toprak and Cigizoglu [27] predicts the highest measured K, 1486.5 m^2/s, as 1050.1 m^2/s, whereas, the present model predicts highest measured K, 892 m^2/s, as 880.7 m^2/s. This suggests that the present ANN model performed better.

Tables 4 gives measured and predicted values of K for testing datasets. Predicted values have been obtained from the present ANN model and from the other important models. As can be seen from Table 4, out of the 20 predictions, eight predictions (40%) by the present ANN model are closest to the measured data, the highest among all models; whereas three predictions by K–F, two each by ANN (T–S), S–C, and S–D models, and one each by Fischer, I–A, and D–S–B are closest to the measured values. The extreme K value of 892 m^2/s is successfully predicted as 880.7 m^2/s by the present ANN model where as it is 763.4 m^2/s by ANN (T–S) model. The Predictions by other models are far from the measured value. A close examination of Table 4 indicates that the present ANN model is giving good predictions for higher values of K. Table 5 shows performance indices of different models for the testing data. As can be observed from Table 5, prediction of the longitudinal dispersion coefficient by the present ANN model is superior to all other models, RMSE being the least and CC the highest for this model. The Performances of ANN (T–S) and I–A can also be called satisfactory as their RMSE$_{\text{s}}$ are comparatively low and CC$_{\text{s}}$ high. The performance of the Fischer model is the least satisfactory. The RMSE improved for all models when extreme values ($K > 100$ m^2/s or $W/H > 50$) are ignored. The most significant improvement is seen in the Fischer model, suggesting the model's inadequacy in predicting K in very wide and large flow rivers. The improvement in the present ANN model is the least, implying that the model captured the pattern well even for extreme cases. If accuracy of a model can be defined as the percentage of DR values falling between -0.175 and 0.175, i.e., predicted values of K lying between 67% and 150% of the measured values, then it can be seen from Table 5 that 65% of the predicted values from the present ANN model are accurate, which is highest among all models. ANN (T–S)'s accuracy is found to be 48%. Fischer model is found to be the least accurate. Another performance indicator of a model, DR range, for the present ANN model is -0.22 to 1.24, which might indicate some skewness of prediction towards positive side. However, if we could ignore the predictions of Bear Creak and John Day (SN 6 and 7, respectively, in Table 4) Rivers, which is acting as outliers for the model, the DR range becomes -0.22 to 0.44 suggesting even distribution. Positive skew ness of Fischer, I–A, and S–D are found to be more pronounced, suggesting over prediction by these models.

Figure 3 shows comparison of the percentage proportion of the predicted values by different models falling into different discrepancy brackets. It shows an almost even distribution for the predictions from the present ANN model. Almost 65% of the predictions by the present model lie between DR -0.2 and 0.2, the highest among all predictions.

4.1 Sensitivity analysis

A sensitivity and error analysis, for prediction of the longitudinal dispersion coefficient, is carried out with the help of five ANN models. Table 6 summarizes the input parameters and

Table 4 Comparison of observed and predicted longitudinal dispersion coefficients

S. no.	Stream	Observed value	ANN (present study)	Longitudinal dispersion coefficient K (m^2/s)							
				ANN (T–S)	Fischer (Eq. 3)	Liu (Eq. 6)	I–A (Eq. 8)	S–C (Eq. 9)	K–F (Eq. 13)	D–S–B (Eq. 10)	S–D (Eq. 14)
1	Antietam Creek, Md.	20.9	44.5	26.8	5.1	7.4	8.6	20.2*	15.2	15.0	13.7
2	Monocacy River, Md.	37.8	22.76	27.1	61.7	90.3	47.7	27.1	7.6	28.2	31.7*
3	Monocacy River, Md.	41.4	46.1*	31.4	74.6	188.1	97.9	23.5	4.2	25.8	33.4
4	Conococheague Creek.	53.3	65.4	43	88.2	66.5	43.0	96.7	58.8*	93.1	78.2
5	Chattahoochee River, Ga	88.9	95.5*	77.6	127.9	168.6	129.9	169.1	82.1	168.6	147.1
6	Bear Creek, Colo.	2.9	50.07	39.2	7.3*	33.6	60.8	52.2	27.1	28.1	39.1
7	Tangipahoa River, La	44	31.4	26.5	142.1	33.2	10.3	39.2*	24.5	28.7	34.7
8	Sabina River, La	308.9	303.41*	346.6	2535.1	477.0	143.9	718.7	512.3	508.4	603.8
9	Sabina River, Tex.	12.8	26.9	21.9	2.0	5.0	5.6*	5.2	2.4	4.6	4.4
10	Mississippi River, La	237.2	325.6*	838	2134.2	691.9	348.3	1854.5	1618.5	1617.7	1327.3
11	Mississippi River, Mo	457.7	504.9*	838	10122.8	2790.4	764.9	1793.0	837.6	1245.1	1834.5
12	Mississippi River, Mo	374.1	420.1*	838	8390.4	2235.4	810.1	3271.5	2220.1	2579.3	2736.0
13	Wind/Big. River, Wyo	41.8	113.9	59.7*	229.6	186.8	103.9	159.9	76.0	156.5	147.0
14	Clinch River, Va	10.7	13.7	26.9	26.4	37.8	27.2	27.6	11.5*	28.4	25.8
15	Clinch River, Va	36.9	48.6*	76.6	52.6	56.2	53.5	139.6	104.1	118.2	97.8
16	Powell River, Tenn.	15.5	26.3	25.3	5.4	23.5	25.8	9.9	2.9	9.9	10.3*
17	John Day River, Ore.	13.9	94.6	45.2	86.4	72.9	46.0	83.3	44.8*	81.7	71.2
18	John Day River, Ore.	65	83.3	77.2	19.3	32.6	45.6	116.7	97.9	71.1*	73.6
19	White River	30.2	37.17	31.7*	229.0	170.0	62.8	54.3	17.7	46.4	65.18
20	Missouri River	892	880.7*	763.4	4119.6	776.0	244.6	1317.3	990.5	950.9	1074.7

Note: *ANN (T–S)* Tayfur and Singh [24], *Fischer* Fischer [4], *Liu* Liu [12], *S–C* Seo and Cheong [21], *D–S–B* Deng et al. [2], *K–F* Kashefipour and Falconer [10], *I–A* Iwasa and Aya [8], *S–D* Sahay RR and Dutta S [19]

* Estimate closest to the measured value

Table 5 Comparison of performance indices of models (for testing data)

Model	RMSE	RMSE ($K > 100\,\mathrm{m}^2/\mathrm{s}$ ignored)	RMSE ($W/H > 50$ ignored)	CC	DR range	Accuracy (%)
ANN (this study)	37.6	32.6	40.6	0.99	−0.22 to 1.24	65
ANN (T–S)	192.9	19.3	182.3	0.83	−0.22 to 1.13	48
Fischer	2972.8	80.5	572.9	0.72	−0.80 to 1.35	10
Liu	679.4	72.3	141.5	0.636	−0.45 to 1.06	15
S–C	812.0	54.3	490.3	0.684	−0.39 to 1.26	20
D–S–B	611.0	47.8	418.4	0.647	−0.44 to 0.98	20
K–F	525.5	27.7	417.3	0.663	−0.99 to 0.97	25
I–A	194.9	33.4	42.4	0.615	−0.63 to 1.32	30
S–D	663.6	41.0	330.4	0.693	−0.46 to 1.30	30

Note: *ANN (T–S)* Tayfur and Singh [24], *Fischer* Fischer [4], *Liu* Liu [12], *S–C* Seo and Cheong [21], *D–S–B* Deng et al. [2], *K–F* Kashefipour and Falconer [10], *I–A* Iwasa and Aya [8], *S–D* Sahay RR and Dutta S [19]

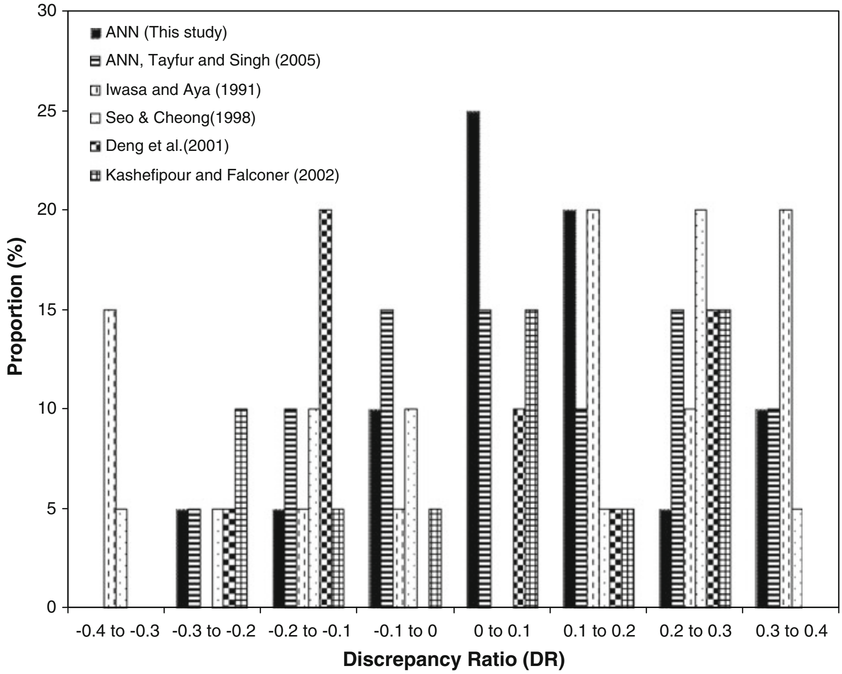

Fig. 3 Comparison of DR values for models

the number of neurons in the hidden layer of models along with their performances. The number of neurons in the hidden layer is obtained by trial and error method. This is the number, which gives the least MSE for the given input datasets.

Model 1 considered stream width (W), average flow depth (H), mean longitudinal flow velocity (U), shear velocity (U_*), and sinuosity (σ) as inputs. The shape parameter (β), which is log value of W/H, is not among input variables, as W and H are expected to capture the

Table 6 ANN models and their performances

ANN models	Inputs	Neurons (in hidden layer)	RMSE (m^2/s)	RMSE (m^2/s) ($K > 100\,m^2/s$ not included)	RMSE (m^2/s) ($W/H > 50\,m^2/s$ not included)	% Accuracy
Model 1	W, H, U, U_*, σ	8	37.6	32.6	40.6	65
Model 2	W, H, U, U_*	8	153.0	50.6	61.6	48.0
Model 3	W, H, U	6	201.9	40.1	98.4	37.0
Model 4	$U/U_*, \sigma$	6	217.7	100.6	157.0	30.0
Model 5	$U/U_*, \beta$	6	209.1	127.0	134.5	26.0

effect of shape variation. Model 2 excludes the sinuosity (σ) from the list of input parameters of Model 1. The objective is to see the effect of exclusion of σ on the accuracy of prediction of K. In Model 3, shear velocity is excluded from the list of input parameters of Model 2, with an objective to see the effect of exclusion U_* on the prediction accuracy of K. In Model 4, the ratio of flow velocity to shear velocity along with the sinuosity is used as input parameter. Model 5 uses U/U_* and β only as inputs. The objective is to check their sufficiency for prediction of dispersion coefficient.

The results of prediction of K from above-mentioned five ANN models can be seen from Table 6. The ANN models, in general, yielded good estimates of measured data. In particular, Model 1 performed best among all models. The RMSE of predictions by this model is the least and accuracy the highest. This suggests that W, H, U, U_*, and σ are important parameter for prediction of K. Ignoring any one from the input dataset would lead to inaccurate prediction. When the extreme values ($K > 100\,m^2/s$ or $W/H > 50$) are excluded from input datasets, then the performance of all models improved. The least performing is Model 5, suggesting its inadequacy for the prediction of K.

5 Conclusion

Some natural physical phenomena like dispersion of pollutants in rivers are very complex, solution of which is found by many simplifications and assumptions. This leads to inadequate prediction. An ANN model, on the other hand, which has shown great promise in a number of multivariate problems in the past, is not bound by such limitations. However, as with any other forecasting model, the success of an ANN model depends largely upon the proper selection of input variables. In this study, a FFBP ANN model is developed to predict the longitudinal dispersion coefficients in rivers. The results of this study show that accuracy in prediction of K improves significantly if input dataset comprises of geometric (stream width, stream sinuosity) and flow (mean longitudinal velocity, mean flow depth, shear velocity) characteristics. For training and testing of the developed ANN model, 71 published datasets of USA rivers are utilized. The performance of the developed ANN model is compared with those of Tayfur and Singh [24], Fischer [4], Liu [12], Iwasa and Aya [8], Seo and Cheong [21], Deng et al. [2], Kashefipour and Falconer [10], and Sahay and Dutta [19]. The Performance indices considered for comparison are RMSE, CC, and DR. Based on the performance indices on testing datasets, the present ANN model is found to be superior to all models considered in this work. In case of wide and high flow rivers, the performance of the present model is even better. Extreme measured values of 892, 457.7, 374.1, and 308.9 m^2/s are successfully predicted as 880.7, 504.9, 420.1, and 303.41 m^2/s, respectively. The ANN model of Tayfur and Singh [24] is found to be less accurate when compared to the present ANN model. This is because they had ignored river sinuosity, which is an important input variable, especially in wide and high flow rivers. Table 5 shows that the models of Tayfur and Singh, Iwasa and Aya, and Kashefipour and Falconer followed the performances of the present ANN model. The performance of Fischer model is found to be the most unsatisfactory, yet his contribution is regarded significant which inspired many investigators in this field.

Appendix I: Optimal weights and biases of the best ANN of the present study

(i) Interconnection weights from hidden neurons to input neurons: [4.6537 −0.55669 0.40181 6.0446 6.0268; −2.4918 −3.6388 −4.4239 3.9328 3.2533; 3.2301 5.6027 4.4843 2.559 1.0168; 3.4226 6.632 −2.2349 −0.92931 11.0702; 13.5052 4.1295 1.5028 −1.5409 −9.1873; −10.6682 3.0621 −4.1874 1.7116 −1.3672; 7.6101 −1.2886 2.2302 1.2365 2.5023; −11.8728 −5.7405 −4.6032 10.6936 −6.0229].

(ii) Interconnection weights from hidden neurons to output neuron: [2.2156 1.9546 −1.1937 8.379 −12.7936 −11.8344 3.6668 −13.9641].

(iii) Bias to neurons in hidden layer: [−10.5285; 7.1862; −10.9455; 24.9784; −2.0838; 0.55192; 7.821; 9.6898].

(iv) Bias to output neuron: [−1.0271].

References

1. ASCE Task Committee (2000) Artificial neural networks in hydrology II. J Hydrol Eng 5(2):124–132
2. Deng ZQ, Singh VP, Bengtsson L (2001) Longitudinal dispersion coefficient in straight rivers. J Hydraul Eng 127(11):919–927
3. Elder JW (1959) The dispersion of a marked fluid in turbulent shear flow. J Fluid Mech 5(4):544–560
4. Fischer BH (1967) The mechanics of dispersion in natural streams. J Hydraul Eng 93(6):187–216
5. Fischer BH, List EJ, Koh RCY, Imberger J, Brooks NH (1979) Mixing in inland and coastal waters. Academic Press, New York, pp 104–138
6. Godfrey RG, Frederick BJ (1970) Stream dispersion at selected sites. US Geological Survey Paper 433-K, Washington, DC
7. Haykin S (1994) Neural networks: a comprehensive foundation. Macmillan College Publishing Co., New York
8. Iwasa Y, Aya S (1991) Predicting longitudinal dispersion coefficient in open channel flows. In: Proceedings of the internation symposium on environmental hydraulics, Hong Kong, pp 505–510
9. Jeon TM, Back KO, Seo IW (2007) Development of an empirical equation for transverse dispersion coefficient in natural streams. Environ Fluid Mech 7(2207):317–329
10. Kashefipour MS, Falconer RA (2002) Longitudinal dispersion coefficients in natural channels. Water Resour Res 36(6):1596–1608
11. Koussis AD, Rodriguez-Mirasol J (1998) Hydraulic estimation of dispersion coefficient in streams. J Hydraul Eng 124(3):317–320
12. Liu H (1977) Predicting dispersion coefficient of streams. J Environ Eng Div 103(1):59–69
13. Mathswork Inc., version 7
14. McQuivey RS, Keefer TN (1974) Simple method for predicting dispersion in streams. J Environ Eng Div 100(4):997–1011
15. Nordin CF, Sabol GV (1974) Empirical data on longitudinal dispersion in rivers. US Geological Survey Water Resource. Invest 20–74, Washington, DC
16. Papadimitrakis I, Orphanos I (2004) Longitudinal dispersion characteristics of rivers and natural streams in Greece. Water Air Soil Pollut 4:289–305

17. Patil S, Li CS, Li X, Li C, Tam BYF, Song CY, Chen YP, Zhang Q (2009) Longitudinal dispersion in wave-current-vegetation flow. Phys Oceanogr 19(1):45–60
18. Piasecki M, Katopodes ND (1999) Identification of stream dispersion coefficient by adjoin sensitivity method. J Hydraul Eng ASCE 125(7):714–724
19. Sahay RR, Dutta S (2009) Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm. Hydrol Res 40(6):544–552
20. Seo IW, Baek KO (2004) Estimation of longitudinal dispersion coefficient in natural streams by artificial neural network. J Hydraul Eng 130(3):227–236
21. Seo IW, Cheong TS (1998) Predicting longitudinal dispersion coefficient in natural streams. J Hydraul Eng 124(1):25–32
22. Shepherd AJ (1997) Second-order methods for neural networks. Springer, New York
23. Sukhodolov AN, Nikora VI, Rowinski PM, Czernuszenko W (1997) A case study of longitudinal dispersion in small lowland rivers. Water Environ Res 69(7):1246–1333
24. Tayfur G, Singh VP (2005) Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. J Hydraul Eng 131(11):991–1000
25. Taylor GI (1953) Dispersion of soluble matter in solvent flowing slowly through a tube. Proc R S Lond A 219:186–203
26. Taylor GI (1954) The dispersion of matter in turbulent flow through a pipe. Proc R S Lond UK A 223: 446–468
27. Toprak ZF, Cigizoglu HK (2008) Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods. Hydrol Process 22:4106–4129