GA-optimized model predicts dispersion coefficient in natural channels

Gokmen Tayfur

ABSTRACT

Models whose parameters were optimized by genetic algorithm (GA) were developed to predict the longitudinal dispersion coefficient in natural channels. Following the existing equations in the literature, ten different linear and nonlinear models were first constructed. The models relate the dispersion coefficient to flow and channel characteristics. The GA model was then employed to find the optimal values of the constructed model parameters by minimizing the mean absolute error function (objective function). The GA model utilized an 80% cross-over rate and 4% mutation rate. It started each computation with a population of 100 chromosomes in the gene pool. For each model, while minimizing the objective function, the values of the model parameters were constrained between [−10, +10] at each iteration. The optimal values of the model parameters were obtained using a calibration set of 54 out of 80 sets of measured data. The minimum error was obtained for the case where the model was a linear equation relating dispersion coefficient to flow discharge. The model performance was then satisfactorily tested against the remaining 26 measured validation datasets. It performed better than the existing equations. It yielded minimum errors of MAE = 21.4 m^2/s (mean absolute error) and RMSE = 28.5 m^2/s (root mean-squares error) and a maximum accuracy rate of 81%.

Key words | cross-over, dispersion coefficient, genetic algorithm, gene pool, mutation, modelling

Gokmen Tayfur
Department of Civil Engineering,
Izmir Institute of Technology,
Gulbahce Kampus,
Urla, Izmir 35340,
Turkey
E-mail: *gokmentayfur@iyte.edu.tr*

NOMENCLATURE

A	cross-sectional area
B	channel width
DR	discrepancy ratio
h	local flow depth
H	cross-sectional average flow depth
K	longitudinal dispersion coefficient
K_m	measured dispersion coefficient
K_p	predicted dispersion coefficient
N	number of observations
U	cross-sectional average flow velocity
u_*	shear velocity
u'	deviation of local depth mean flow velocity from cross-sectional mean
y	coordinate in the lateral direction
$\alpha, \beta, \varepsilon, \eta, \delta$	parameters

doi: 10.2166/nh.2009.010

INTRODUCTION

The longitudinal dispersion coefficient is a fundamental parameter in hydraulic modelling of river pollution. It is a measure of mixing of the intensity of the pollutants in natural streams. Hence, it has been extensively investigated (Elder 1959; Sooky 1969; McQuivery & Keefer 1976; Sukhodolov *et al.* 1997). Taylor (1953, 1954) first introduced the longitudinal dispersion coefficient as a measure of the one-dimensional dispersion process and Fischer *et al.* (1979) developed the integral expression:

$$K = \frac{1}{A}\int_0^B hu' \int_0^y \frac{1}{\varepsilon_t h}\int_0^y hu' \,dy\,dy\,dy \tag{1}$$

where K is longitudinal dispersion coefficient; A is

cross-sectional area; B is channel width; h is local flow depth; u' is deviation of local depth mean flow velocity from cross-sectional mean; y is coordinate in the lateral direction; and ε_t is local transverse mixing coefficient.

Due to the requirement for detailed transverse profiles of velocity and cross-sectional geometry, it is rather difficult to use Equation (1). As a result, investigators have proposed empirical equations based on experimental and field data for predicting the dispersion coefficient (e.g. Fischer *et al.* 1979; Seo & Cheong 1998; Kashefipour & Falconer 2002). The proposed empirical equations can be expressed in a general form:

$$K = \alpha\left(\frac{U}{u_*}\right)^{\beta}\left(\frac{B}{H}\right)^{\eta} \tag{2}$$

where H is cross-sectional average flow depth; u_* is shear velocity; U is cross-sectional average flow velocity; and α, β, and η are coefficients which are mostly found through regression analysis.

Deng *et al.* (2001, 2002) developed theoretically based models from Equation (1). Their first model is semi-theoretical and has the form of Equation (2). It not only includes the conventional parameters of the hydraulic variables (B/H) and the friction term (U/u_*), but also the effects of the transverse mixing (ε_{to}). Their last model is fully theoretical which has a general applicability for a wide range of field conditions. However, this model has a major drawback in terms of its complexity coming from an application of approximation methods for triple numerical integration with a set of regression equations (Rowinski *et al.* 2005).

Recently, the artificial neural networks have been employed in the prediction of the dispersion coefficient using flow and channel geometric characteristics (Tayfur & Singh 2005; Rowinski *et al.* 2005; Tayfur 2006). ANNs have an ability to capture relationships from given patterns and this ability has enabled them to be employed in the solution of large-scale complex problems. However, ANNs are black box models that do not reveal any physical relations between the input and the output variables of the system. It is therefore difficult to have a good insight of the physical process. Furthermore, although ANNs are good interpolators, they mostly lack extrapolation capability. They generally perform poorly for the cases for which they are not trained (Tayfur *et al.* 2007).

This study proposes empirical equations following the literature and finds the optimal values of the coefficients of the formulations using the genetic algorithm (GA) method which has recently found wide application in water resources engineering (Liong *et al.* 1995; Guan & Aral 1998; Sen & Oztopal 2001; Jain *et al.* 2004; Guan & Aral 2005; Singh & Datta 2006; Aytek & Kisi 2008; Hejazi *et al.* 2008; Tayfur & Moramarco 2008). The different formulations cover possible combinations relating flow and channel characteristics to the dispersion coefficient.

This paper is organized such that the following section summarizes proposed formulations, followed by a brief background on GA. Application of genetic algorithms to find the optimal values of the parameters of the proposed models is then presented. Afterwards, the comparative performance analysis of the best performing constructed model against the commonly employed existing equations is presented. The summary and conclusions follow the analysis of the results.

PROPOSED MODELS

As seen in Equation (2) as well as in existing empirical equations in the literature (Seo & Cheong 1998), the longitudinal dispersion coefficient (K) is predicted from flow variables and channel geometric characteristics (U, u_*, H, B, B/H, U/u_*, Q). Following Equation (2), this study further proposed:

$$K = \alpha\left(\frac{U^{\beta}}{u_*^{\varepsilon}}\right)\left(\frac{B^{\eta}}{H^{\delta}}\right) \tag{3}$$

$$K = \alpha\left[\frac{BU}{Hu_*}\right]^{\beta} \tag{4}$$

Deng *et al.* (2002), Rowinski *et al.* (2005) and Tayfur (2006) predicted K using only flow velocity and channel width data. Following those studies, we propose:

$$K = \alpha(UB)^{\beta} \tag{5}$$

$$K = \alpha(U^{\beta}B^{\varepsilon}) \tag{6}$$

Tayfur & Singh (2005) and Tayfur (2006) predicted K from flow discharge (Q) and flow shear velocity (u_*) data.

Following those studies, we propose

$$K = \alpha\left(\frac{Q}{u_*}\right)^{\beta} \tag{7}$$

$$K = \alpha\frac{Q^{\beta}}{u_*^{\varepsilon}} \tag{8}$$

$$K = \alpha\left[\frac{B^{\beta}U^{\varepsilon}H^{\eta}}{u_*^{\delta}}\right] \tag{9}$$

Tayfur (2006) predicted K from only the flow discharge data. Hence, we propose:

$$K = \alpha(Q)^{\beta} \tag{10}$$

$$K = \alpha Q + \beta \tag{11}$$

Equations (2–11) might cover all the possible combinations of flow and channel characteristics for finding the dispersion coefficient. This study found the optimal values of the coefficients of Equations (2–11) by the genetic algorithm method.

GENETIC ALGORITHM (GA)

Genetic algorithm is a nonlinear search and optimization method inspired by biological processes of natural selection and the survival of the fittest. Basic units of GA consist of *bit*, *gene*, *chromosome* and *gene pool*. *Gene* consisting of bits (0 and 1) represents a model parameter (or a decision variable) to be optimized. The combination of genes forms the *chromosome*, each of which is a possible solution for each variable. Finally, a set of chromosomes forms the *gene pool*.

The main GA operations basically consist of generation of initial gene pool, evaluation of fitness for each chromosome, selection, cross-over and mutation. An initial population of chromosomes can be randomly generated by, for example, a uniform distribution or a normal distribution. Fitness of each chromosome can be obtained as (Sen 2004):

$$F(C_i) = \frac{f(C_i)}{\sum f(C_i)} \tag{12}$$

where C_i is chromosome i; $F(C_i)$ is fitness value of chromosome that is the percentage of variable in the pool; and $f(C_i)$ is the value of objective function evaluated for chromosome i.

Selection can be performed randomly by, for example, a roulette wheel (Sen 2004) or by ranking the chromosomes according to their fitness from the fittest to weakest. The fittest ones are then copied from the weakest ones. By *cross-over*, new individuals are produced by changing the genes of the chromosomes. The last operation in GA is the *mutation* where a particular bit (bits) is reversed (i.e. 1 to 0 or 0 to 1). In a GA search, this is the perturbation that allows the GA to seek out new and novel solutions. Figure 1 is an example demonstrating that the value of 153 goes to 57 after cross-over and then to 249 after mutation, scanning a large area of the solution domain. The details of GA can be obtained from e.g. Goldberg (1989), Sen (2004) and Eiben & Smith (2007).

GA application

In order to obtain optimal values of the parameters (α, η, ε, δ) of the above proposed Equations (Equations (2–11)) by the GA and test the performance of the equations, this study employed 80 sets of measured field data (see Appendix) from the literature (Seo & Cheong 1998; Deng *et al.* 2001, 2002; Kashefipour & Falconer 2002). The statistics (mean, standard deviation, minimum and maximum) for each variable are also summarized in the same Appendix.

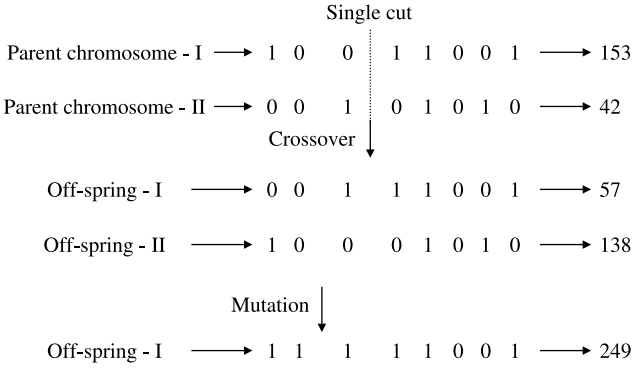

Figure 1 | Example of a cross-over and mutation operations.

The datasets in the Appendix were separated into two groups as calibration (54 sets) and validation sets (26 sets). The calibration set was used for calibrating the constructed models by the GA while the other set was used for verifying the developed models. Although the data for calibration and validation sets were chosen randomly, special attention was paid so that the statistics of the sets would have comparable orders of magnitude. Such precautions avoid bias in the model predictions. Table 1 summarizes the statistics for both the sets. As seen in Table 1, the statistics are in the same order of magnitude.

In order to find the optimal values of the model parameters in each constructed model (Equations (2–11)), 10 different runs with 5,000 iterations were performed. Note that the GA model can produce different values for model parameters at each run due to the initial random values assigned for each model parameter and its major operations of cross-over and mutation. The program converges faster and no vital change takes place other than slight accuracy improvements after 3,000 iterations. Initially, parameters were assigned random values in the [0, 2] range. The GA model employed 100 chromosomes in the initial gene pool, 80% cross-over rate and 4% mutation rate. During the GA search process, the range for each parameter in each model was constrained in $[-10, +10]$ at each iteration. The trial version of evolver GA solver for Microsoft Excel (Palisade Corporation 2001) was employed in this study. The algorithm employs the *Recipe Solving Method* to minimize/maximize objective function under specified constraints (Palisade Corporation 2001). It is very easy to construct the GA on Excel; it takes a very short time (of order seconds), to run the program for thousands of iterations.

Optimal parameter values were found by minimizing the mean absolute error objective function (MAE) having the following form:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|K_{\mathrm{pred},i} - K_{\mathrm{meas},i}\right| \tag{13}$$

where N is the number of observations; K_{pred} is the predicted dispersion coefficient and K_{meas} is the measured dispersion coefficient.

The MAE, illustrating the possible maximum deviation, is one of the commonly employed error functions in the literature (Chang *et al.* 2005). According to Taji *et al.* (1999), the *absolute* error may sometimes be better than the square error to minimize the deviation. In fact, the absolute error function has the advantage that it is less influenced by anomalous data than the square error function (Taji *et al.* 1999).

The calibration results for each model (Equations (2–11)) are summarized in Tables 2–11. The minimum error is obtained for the model expressed by Equation (11) with MAE = 26.7 m^2/s (Table 11). As also seen in Table 11, all the runs for this model produced errors less than 30 m^2/s. Equations (3), (5), (6) and (8–10) had comparable performances with errors around MAE = 31 m^2/s (see Tables 3, 5, 6, 8–10). Although these models produced minimum errors, most of the runs for each model produced large errors, around MAE = 40 m^2/s. Equations (2) and (4) yielded

Table 1 | Statistics for calibration and verification datasets ($\bar{X}$: mean; $X_{\min}$: minimum value; $X_{\max}$: maximum value; SD: standard deviation)

	B (m)	*H* (m)	*U* (m/s)	*u*∗ (m/s)	*B/H* (–)	*U/u*∗ (m^3/s)	*Q* (m^2/s)	*K* (–)
Calibration set								
$\bar{X}$	52.4	1.12	0.47	0.08	49.9	7.13	63.5	75.3
$X_{\min}$	12.2	0.22	0.034	0.002	15.6	1.20	1.0	1.9
$X_{\max}$	253.6	3.56	1.74	0.27	156.5	19.06	915.9	837.0
SD	52.2	0.81	0.332	0.074	31.2	3.84	147.2	141.8
Validation set								
$\bar{X}$	58.1	1.36	0.56	0.09	50.2	6.87	91.7	91.6
$X_{\min}$	11.9	0.40	0.13	0.02	16.1	2.7	0.9	2.9
$X_{\max}$	161.5	3.96	1.53	0.55	131.0	19.63	937.4	892.0
SD	46.4	1.07	0.36	0.06	31.9	4.59	189.7	176.0

highest errors with MAE = 57.2 and 56.2 m^2/s, respectively (see Tables 2 and 4). Equation 7 had MAE = 43.4 m^2/s (Table 7). Tables 2–11 also present the optimal values of the model parameters for each corresponding model. The optimal parameter values that resulted in the minimum MAE for the model represented by Equation (11) are $\alpha = 0.91$ and $\beta = 9.94$. This model can then be expressed as:

$$K = 0.91Q + 9.94 \quad (14)$$

Equation (14) implies that K can be predicted from Q data using the optimal parameter values predicted by the GA model. This is in agreement with Tayfur & Singh (2005) and Tayfur (2006).

The performance of these constructed models (Equations (2–11)) was also tested against the validation data set (see Appendix). For this purpose, for each equation,

Table 2 | MAE and optimal model parameters for Equation (2)

	α	β	η	MAE
Run 1	0.01	0.89	1.54	63.2
Run 2	0.75	1.53	0.25	60.2
Run 3	0.25	1.00	0.78	57.8
Run 4	1.59	1.00	0.31	57.2
Run 5	1.00	−0.17	0.94	61.3
Run 6	0.25	1.49	0.39	60.6
Run 7	0.25	1.74	0.25	61.6
Run 8	1.21	−0.41	1.0	63.7
Run 9	7.72	0.03	0.33	59.5
Run 10	3.47	0.75	0.25	57.8

Table 3 | MAE and optimal model parameters for Equation (3)

	α	β	η	ε	δ	MAE
Run 1	0.98	2.64	1.00	0.70	0.83	40.2
Run 2	0.87	1.94	1.18	0.25	0.46	32.4
Run 3	0.25	6.47	0.75	1.25	2.00	60.1
Run 4	1.18	1.69	1.43	−0.37	0.42	40.5
Run 5	0.25	3.14	1.00	1.00	0.35	41.5
Run 6	0.25	1.00	0.25	0.48	−4.26	48.9
Run 7	4.94	2.19	0.48	0.73	−0.11	34.8
Run 8	1.00	1.34	0.55	0.73	−1.33	33.0
Run 9	1.05	1.54	0.16	0.99	−2.42	41.6
Run 10	0.25	6.34	0.41	1.25	0.09	59.9

Table 4 | MAE and optimal model parameters for Equation (4)

	α	β	MAE
Run 1	9.83	0.25	59.7
Run 2	9.81	0.10	62.3
Run 3	8.47	0.23	58.9
Run 4	2.33	0.47	57.1
Run 5	1.75	0.53	56.6
Run 6	1.67	0.54	56.5
Run 7	1.75	0.52	56.6
Run 8	0.59	0.71	56.2
Run 9	6.01	0.29	58.5
Run 10	1.68	0.55	56.4

Table 5 | MAE and optimal model parameters for Equation (5)

	α	β	MAE
Run 1	2.37	1.00	34.7
Run 2	0.50	1.34	31.5
Run 3	0.50	1.35	33.4
Run 4	8.05	0.66	41.1
Run 5	2.00	1.06	33.6
Run 6	1.61	1.09	33.3
Run 7	5.03	0.77	37.9
Run 8	3.65	0.87	36.6
Run 9	9.80	0.50	45.9
Run 10	8.60	0.61	41.7

Table 6 | MAE and optimal model parameters for Equation (6)

	α	β	ε	MAE
Run 1	1.32	1.45	1.14	30.9
Run 2	7.15	1.50	0.75	34.4
Run 3	6.60	1.63	0.75	35.6
Run 4	2.85	1.14	1.00	33.4
Run 5	9.52	0.10	0.34	54.2
Run 6	0.76	1.50	1.25	30.6
Run 7	0.87	1.27	1.25	34.2
Run 8	8.07	1.26	0.73	35.6
Run 9	9.75	0.63	0.50	45.6
Run 10	1.25	1.17	1.17	32.4

Table 7 | MAE and optimal model parameters for Equation (7)

	α	β	MAE
Run 1	2.80	0.50	45.4
Run 2	0.36	0.78	45.3
Run 3	0.50	0.76	43.4
Run 4	1.84	0.56	44.4
Run 5	1.27	0.61	44.1
Run 6	1.50	0.60	44.1
Run 7	5.87	0.36	48.3
Run 8	9.92	0.20	54.9
Run 9	9.99	0.25	52.6
Run 10	8.19	0.30	50.7

Table 8 | MAE and optimal model parameters for Equation (8)

	α	β	ε	MAE
Run 1	1.46	0.74	0.44	33.1
Run 2	6.03	0.50	0.25	40.0
Run 3	5.97	0.62	0.11	34.3
Run 4	9.00	0.38	0.20	45.1
Run 5	4.78	0.58	0.21	35.2
Run 6	2.15	1.00	−0.32	36.8
Run 7	0.49	0.96	0.36	29.2
Run 8	8.98	0.35	0.20	47.1
Run 9	2.14	0.50	0.60	38.1
Run 10	8.45	0.52	0.12	37.1

Table 9 | MAE and optimal model parameters for Equation (9)

	α	β	ε	η	δ	MAE
Run 1	0.24	0.23	3.92	0.87	1.94	40.1
Run 2	0.25	1.31	1.33	0.25	0.25	32.2
Run 3	0.46	1.50	−1.18	1.07	−1.24	57.9
Run 4	0.50	1.50	−1.19	0.55	−1.05	56.0
Run 5	1.64	1.50	−0.01	0.21	−1.02	52.7
Run 6	0.25	1.13	1.36	0.66	0.40	31.3
Run 7	9.72	0.15	0.25	1.11	0.43	51.3
Run 8	9.95	0.25	0.99	1.53	0.34	31.7
Run 9	1.05	0.47	1.76	1.00	1.00	33.0
Run 10	1.00	1.25	2.38	−1.15	0.19	39.8

Table 10 | MAE and optimal model parameters for Equation (10)

	α	β	MAE
Run 1	9.91	0.46	40.8
Run 2	2.47	0.85	29.7
Run 3	8.11	0.59	34.8
Run 4	1.55	0.94	31.9
Run 5	0.64	1.06	32.9
Run 6	1.09	0.97	31.2
Run 7	6.52	0.67	32.8
Run 8	0.25	1.25	40.8
Run 9	9.91	0.50	37.4
Run 10	9.88	0.28	55.2

Table 11 | MAE and optimal model parameters for Equation (11)

	α	β	MAE
Run 1	0.74	9.64	29.8
Run 2	0.95	9.88	27.4
Run 3	0.82	9.79	28.0
Run 4	0.92	9.95	26.9
Run 5	0.91	9.94	26.7
Run 6	0.92	9.82	26.9
Run 7	0.87	9.59	27.2
Run 8	0.89	9.94	26.8
Run 9	0.96	9.92	27.6
Run 10	0.81	9.84	28.2

Table 12 | MAE, RMSE and accuracy

Equation	MAE (m^2/s)	RMSE (m^2/s)	Accuracy (%)
(2)	72.1	163.5	50
(3)	52.2	86.6	46
(4)	76.1	174.6	58
(5)	38.1	61.6	58
(6)	39.6	62.7	50
(7)	34.7	66.6	62
(8)	26.1	39.2	62
(9)	33.9	55.1	50
(10)	27.6	37.3	69
(11)	21.4	28.5	81
Mean	42.2	77.6	58.6
Std. dev.	18.9	51.1	10.6

Table 13 | Prediction results of measured K (m^2/s) data by the GA-optimized models

River	Measured K (m^2/s)	Fischer *et al.* (1979)	Seo & Cheong (1998)	Deng *et al.* (2001)	Kashefipour & Falconer (2002)	Tayfur (2006)	GA Equation (11)
Antietam Creek, MD	20.9	5.1	20.2	15.3	15.2	24.8	13.0
Monocacy River, MD	37.8	61.6	27.2	29.0	7.6	28.1	16.3
Monocacy River, MD	41.4	74.7	23.6	26.6	4.2	31.3	19.6
Conococh.Creek, MD	53.3	88.5	96.8	95.5	58.7	49.5	37.8
Cattahooc. River, GA	88.9	128.0	169.3	173	82.1	120.7	109.2
Cattahooc. River, GA	166.9	109.7	148.0	146.6	68.8	127.5	116.0
Bear Creek, CO	2.9	7.3	52.3	28.7	27.1	35.3	23.6
Tangipah. River, LA	44.0	142.3	39.3	29.5	24.5	25.4	13.7
Red River, LA	130.5	101.5	132.9	134.4	58.8	190.0	178.7
Sabina River, LA	308.9	2535.5	719.3	522.4	512.2	379.3	368.7
Sabina River, TX	12.8	2.0	5.2	4.7	2.4	22.5	10.8
Wind/Big River, WY	41.8	229.3	160.0	160.7	75.9	73.9	62.3
Powell River, TN	9.5	6.9	12.5	12.7	4.2	25.9	14.1
Clinch River, VA	10.7	26.3	27.6	29.2	11.5	27.2	15.5
Clinch River, VA	36.9	52.7	139.7	121.0	104.1	98.4	86.9
Powell River, TN	15.5	5.4	9.9	10.1	2.9	25.5	13.8
John Day River, OR	13.9	86.3	83.3	83.8	44.8	35.0	23.2
John Day River, OR	65.0	19.4	116.8	72.6	97.9	84.4	72.8
Yadkin River, NC	260.1	66.0	277.0	177.2	183.7	211.3	200.1
White River	30.2	232.9	55.8	49.8	17.4	34.2	22.5
Missouri River	892.0	4115.8	1318.4	976.9	990.3	871.9	863.0
Clinch River, TN	46.5	87.6	171.3	157.7	129.3	101.5	90.0
Antietam Creek	16.3	22.0	27.7	27.9	14.8	25.7	13.9
Monocacy River	13.9	66.2	28.6	29	9.6	26.5	14.8
Elkhorn River	20.9	312.2	60	48.6	20.5	30.6	18.9
Muddy Creek	32.5	7.5	35.4	24.9	27.7	31.2	19.5
Mean	91.6	330.5	152.2	122.6	99.8	105.3	93.8
Std dev	176.0	895.2	271.7	199.6	204.2	172.6	173.2

the corresponding optimal parameter values obtained by the GA using the calibration dataset were employed (see Tables 2–11). These equations were applied to predict the 26 measured dispersion coefficients in the validation dataset. Table 12 summarizes the computed MAE, RMSE and the accuracy, together with mean and standard deviation values, for each equation. The accuracy of each model may be categorized by the number of discrepancy ratio (DR), defined

$$DR = \log\left(\frac{K_{\text{pred}}}{K_{\text{meas}}}\right)$$

and valued between -0.3 and 0.3 relative to the total number of data values (Kashefipour & Falconer 2002). According to Table 12, Equation (11) (Equation (14)) produced minimum errors (MAE = 21.4 m^2/s, RMSE = 28.5 m^2/s) and the

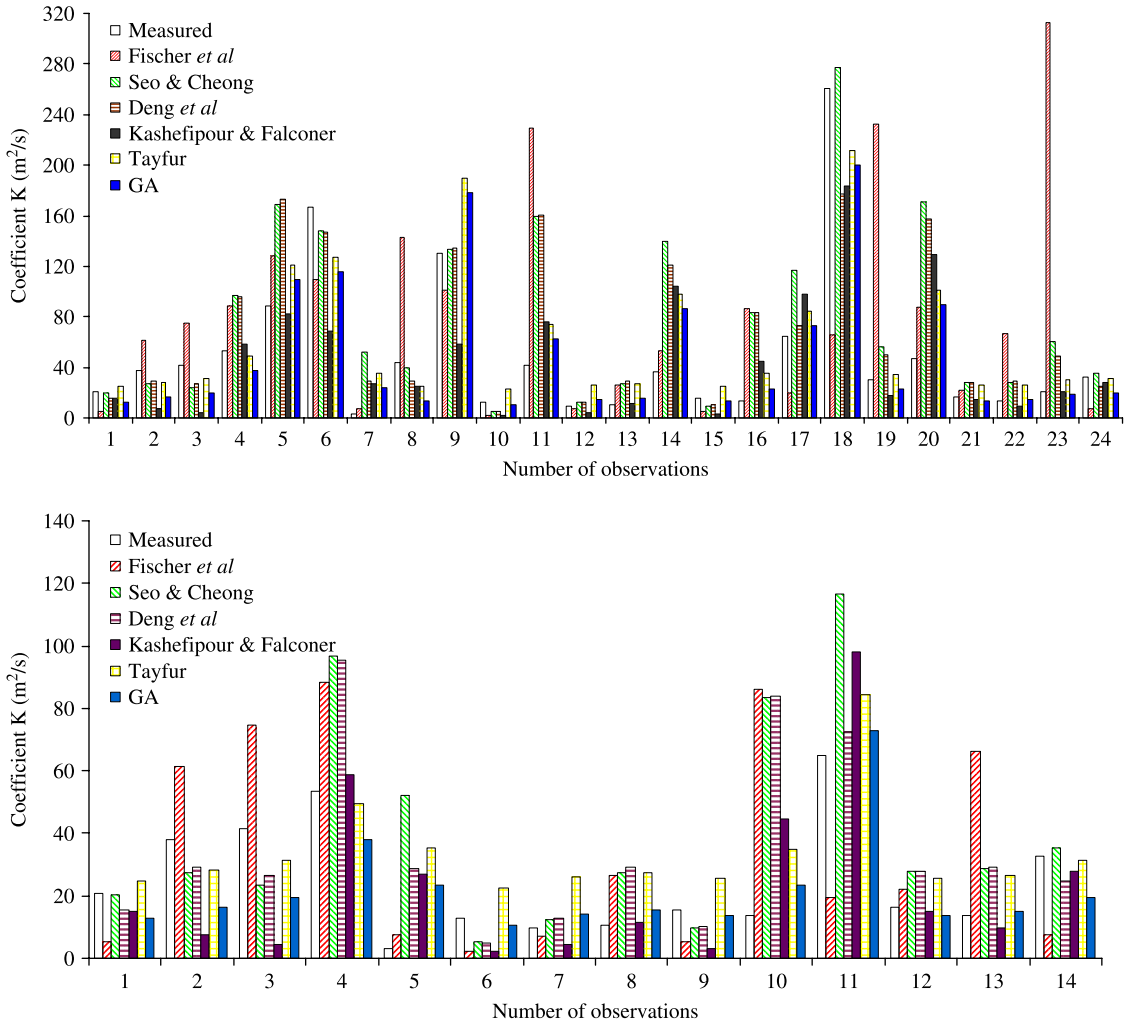

Figure 2 | Comparison of models in predicting measured dispersion coefficient.

maximum accuracy of 81% for the validation dataset. Equation (11) is followed by Equation (10) with 69% accuracy. Most of the other equations showed comparable accuracy around 60% while Equations (2) and (4) produced the largest MAE and RMSE values.

COMPARATIVE ANALYSIS AND DISCUSSION

As presented above, among the proposed Equations (2–11), Equation (11) produced minimum errors and maximum accuracy for both the calibration and validation datasets (Tables 11 and 12). Its performance was therefore also tested against the existing models. Table 13 presents the predictions of the dispersion coefficients of the validation dataset, together with mean and standard deviation values by Equation (14) and the existing equations. According to Table 13, 12 of 26 measured data were closely predicted by the GA-optimized model (Equation (14)). It predicted 46% of the measured data better, which is the highest of all. The comparison results are also presented in Figure 2 as a bar chart and in Figure 3 as scatter diagrams. As seen in these figures, the GA-optimized model (Equation (14)) performs, in general, satisfactorily in predicting low and as well as high values of the dispersion coefficient. In those figures, the extreme values, namely #10 and #21 in Table 13, were not shown for the sake of clarity.

In order to objectively evaluate the model performances, MAE, RMSE and DR values were also computed for

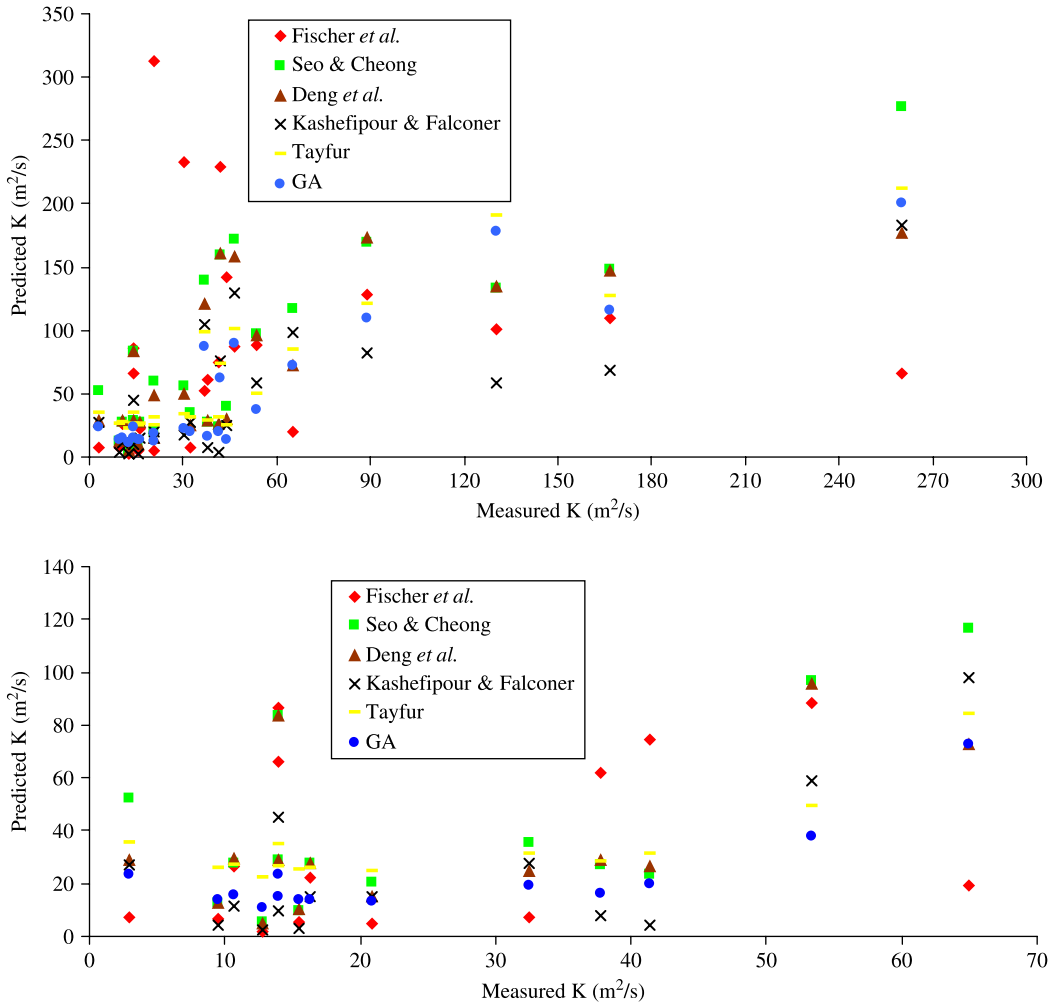

Figure 3 | Scatter diagram showing model predictions of measured dispersion coefficient.

each model. Table 14 shows the computed MAE, RMSE and accuracy for each model. As seen in Table 14, the GA-optimized model has the lowest errors with MAE = 21.4 m^2/s and RMSE = 28.5 m^2/s while the model developed by Fischer *et al.* (1979) has the highest error values with MAE = 267.7 m^2/s and RMSE = 774.1 m^2/s. The model developed by Tayfur (2006) follows the GA-optimized model with comparable errors of MAE = 24.1 m^2/s and RMSE = 31.1 m^2/s. The models developed by Kashefipour & Falconer (2002) and Deng *et al.* (2001) have comparable performances (Table 14). As seen in Table 14, the accuracy of the GA-optimized model (Equation (14)) is 81%, which is the highest of all.

Table 14 | MAE, RMSE, accuracy and *t*-values ($\alpha = 0.05$ level of significance and degree of freedom of 25; the critical value for Student's *t*-distribution is 1.708)

	GA	Fischer *et al.* (1979)	Seo & Cheong (1998)	Deng *et al.* (2001)	Kashefipour & Falconer (2002)	Tayfur (2006)
MAE (m^2/s)	21.4	267.7	64.5	42.7	37.6	24.1
RMSE (m^2/s)	28.5	774.1	126.0	62.2	58.8	31.1
Accuracy (%)	81	42	62	65	58	76
t-values	1.115	1.539	1.795	1.695	1.091	1.463

The performances of the models were finally subjected to the t-test. Note that the sample size is 26 (the number of verification data sets in Table 13) and therefore the degree of freedom (DF) is 25 (Mason *et al.* 1998). For $\alpha = 0.05$ level of significance and DF = 25, the critical value of t-distribution for one-tailed test is 1.708 (Mason *et al.* 1998). Table 14 presents the computed t-values for each model in Table 13. As seen, with the exception of the model of Seo and Cheong, other models produced t-values below the critical value of 1.708. The GA-based optimized model (Equation (14)) and the model of Kashefipour and Falconer produced the lowest t-values. In short, all the models (except for the model of Seo & Cheong) passed the t-test.

Above qualitative results imply that the GA-optimized model, which is very simple to develop and implement, has superiority over the existing models; it can also be employed for predictive purposes. It should be noted, however, that Equation (14) is derived using data from widely seen natural streams whose width is mostly wide (on average 54 m) having normal flow depth (on average 1.2 m), flow velocity (on average 0.50 m/s) and flow discharge (on average 69 m^3/s) (see Appendix).

On the other hand, mountainous streams have generally higher bed slopes and narrower cross-sections with lower flow depths but faster flow velocities. Hence, for the same flow discharge, the intensity of the dispersion process is expected to be different in mountainous and commonly seen natural streams. On the other hand, according to Equation (14), one would predict the same K value for the same Q value from two different streams: a fast-flowing mountainous stream and a normally flowing natural stream. Therefore, one needs to be cautious when employing Equation (14) to predict K in fast-flowing mountainous streams. Furthermore, one also needs to pay attention when using Equation (14) to predict K in a stream whose discharge rate is very low. Equation (14) would over-predict K for a very low Q value (i.e. when $Q = 0.0$, Equation (14) yields $K = 9.94$ m^2/s).

SUMMARY AND CONCLUSIONS

Several linear and nonlinear models that are consistent with the literature were first constructed and a GA model was employed to predict the optimal values of the parameters of the models. Eighty sets of data from 30 natural streams were separated into two groups as calibration and validation sets. The calibration set consisting of 56 data samples was used for calibrating the models. The minimum error was obtained for the case where the linear model expressed by Equation (11) (Equation (14)) predicts the coefficient (K) from flow discharge (Q). The GA-optimized constructed models (Equations (2–11)) were also tested against the validation set, consisting of 26 data samples. The GA-optimized model expressed by Equation (11) (Equation (14)) also produced minimum errors and maximum accuracy among the constructed models.

The performance of the GA-optimized model (Equation (14)) was then tested against the existing equations for the validation dataset. The results revealed that the GA-optimized model predicts the measured data satisfactorily with minimum MAE, RMSE and maximum accuracy.

The satisfactory performance of the model against measured data and as well as the existing linear and nonlinear equations revealed that the GA-optimized model can be a promising modelling tool for predictive purposes. Thus the GA-optimized model developed in this study can be an alternative to the existing equations for predicting the dispersion coefficient in natural streams.

It should however be noted that the model was developed using data from widely seen natural streams presented in the Appendix and it may therefore have a limited predictive capacity for K of fast-flowing mountainous streams. It may also have a poor predictive capacity for K of a stream having very low flow discharge rate.

REFERENCES

Aytek, A. & Kisi, O. 2008 A genetic programming approach to suspended sediment modelling. *J. Hydrol.* **351**(3–4), 288–298.

Chang, C. L., Lo, S. L. & Yu, S. L. 2005 Applying fuzzy theory and genetic algorithm to interpolate precipitation. *J. Hydrol.* **314**(1–4), 92–104.

Deng, Z. Q., Singh, V. P. & Bengtsson, L. 2001 Longitudinal dispersion coefficient in straight rivers. *J. Hydraul. Eng.* **127**(11), 919–926.

Deng, Z. Q., Bengtsson, L., Singh, V. P. & Adrian, D. D. 2002 Longitudianl dispersion coefficient in single-channel streams. *J. Hydraul. Eng.* **128**(10), 901–916.

Eiben, A. E. & Smith, J. E. 2007 *Introduction to Evolutionary Computing*. Springer, Natural Computing Series.

Elder, J. W. 1959 The dispersion of a marked fluid in turbulent shear flow. *J. Fluid Mech.* **5**(4), 544–560.

Fischer, H. B., List, E. J., Koh, R. C. Y., Imberger, J. & Brooks, N. H. 1979 *Mixing in Inland and Coastal Waters*. Academic, New York, pp. 104–138.

Goldberg, D. E. 1989 *Genetic algorithms for search, optimization, and machine learning*. Addison-Wesley, USA.

Guan, J. & Aral, M. M. 1998 Progressive genetic algorithm for solution of optimization problems with nonlinear equality and inequality constraints. *J. Appl. Math. Model.* **23**, 329–343.

Guan, J. & Aral, M. M. 2005 Remediation system design with multiple uncertain parameters using fuzzy sets and genetic algorithm. *J. Hydrol. Eng.* **10**(5), 386–394.

Hejazi, M. I., Cai, X. M. & Borah, D. K. 2008 Calibrating a watershed simulation model involving human interference: an application of multi-objective genetic algorithms. *J. Hydroinformatics* **10**(1), 97–111.

Jain, A., Bhattacharjya, R. K. & Sanaga, S. 2004 Optimal design of composite channels using genetic algorithm. *J. Irrigation Drainage Eng.* **130**(4), 286–295.

Kashefipour, M. S. & Falconer, R. A. 2002 Longitudinal dispersion coefficients in natural channels. *Water Res.* **36**(6), 1596–1608.

Liong, S. Y., Chan, W. T. & ShreeRam, J. 1995 Peak flow forecasting with genetic algorithm and SWMM. *J. Hydraul. Eng.* **121**(8), 613–617.

Mason, D. M., Lind, D. A. & Marchal, W. G. 1998 *Statistics*. Duxbury Press, International Thomson Publishing Company, USA.

McQuivery, R. S. & Keefer, T. N. 1976 Convective model of longitudinal dispersion. *J. Hydraul. Div.* **102**(10), 1409–1424.

Palisade Corporation 2001 *Evolver, the Genetic Algorithm Solver for Microsoft Excel*. Newfield, New York, USA.

Rowinski, P. M., Piotrowski, A. & Napiorkowski, J. J. 2005 Are artificial neural network techniques relevant for the estimation of longitudinal dispersion coefficient in rivers? *Hydrol. Sci. J.* **50**(1), 175–187.

Sen, Z. 2004 Genetic Algorithm and Optimization Methods. Su Vakfi Yayinlari, Istanbul. (Turkish), ISBN: 975-6455-12-8.

Sen, Z. & Oztopal, A. 2001 Genetic algorithms for the classification and prediction of precipitation occurrence. *Hydrol. Sci. J.* **46**(2), 255–267.

Seo, I. W. & Cheong, T. S. 1998 Predicting longitudinal dispersion coefficient in natural streams. *J. Hydraul. Eng.* **124**(1), 25–32.

Singh, R. M. & Datta, B. 2006 Identification of Groundwater Pollution Sources Using GA-based Linked Simulation Optimization Model. *J. Hydrol. Eng.* **11**(2), 101–109.

Sooky, A. A. 1969 Longitudinal dispersion in open channels. *J. Hydraul. Eng.* **95**(4), 1327–1346.

Sukhodolov, A. N., Nikora, V. I., Rowinsky, P. M. & Czernuszenko, W. 1997 A case study of longitudinal dispersion in small lowland rivers. *Water Environ. Res.* **69**(7), 1246–1333.

Taji, K., Miyake, T. & Tamura, H. 1999 On error back propagation algorithm using absolute error function. *International Conference on Systems, Man, and Cybernetics, IEEE SMC'99*, Vol. 5, 401–406.

Tayfur, G. 2006 Fuzzy, ANN, and regression models to predict longitudinal dispersion coefficient in natural streams. *Nordic Hydrol.* **37**(2), 143–164.

Tayfur, G. & Singh, V. P. 2005 Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. *J. Hydraul. Eng.* **131**(11), 991–1000.

Tayfur, G. & Moramarco, T. 2008 Predicting hourly-based flow discharge hydrographs from level data using genetic algorithms. *J. Hydrol.* **352**(1–2), 77–93.

Tayfur, G., Moramarco, T. & Singh, V. P. 2007 Predicting and forecasting flow discharge at sites receiving significant lateral inflow. *Hydrol. Processes* **21**, 1848–1859.

Taylor, G. I. 1953 Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc. R. Sci. A* **219**, 186–203.

Taylor, G. I. 1954 The dispersion of matter in turbulent flow through a pipe. *Proc. R. Sci. A* **223**, 446–468.

First received 16 July 2007; accepted in revised form 3 November 2008

APPENDIX A

Appendix A: | Experimental data of channel characteristics, flow variables and the dispersion coefficient in natural streams (Seo & Cheong 1998; Deng *et al.* 2001, 2002; Kashefipour & Falconer 2002)

No	Stream	B (m)	H (m)	U (m/s)	u_* (m/s)	B/H (−)	U/u_* (−)	Q (m^3/s)	K (m^2/s)
Calibration Set									
1	Antietam Creek, MD	12.8	0.3	0.42	0.057	42.7	7.37	1.6	17.5
2	Antietam Creek, MD	21	0.48	0.62	0.069	43.8	8.99	6.2	25.9
3	Monocacy River, MD	97.5	1.15	0.32	0.058	84.8	5.52	35.9	119.8
4	Conococheague Creek, MD	42.2	0.69	0.23	0.064	61.2	3.59	6.7	40.8
5	Conococheague Creek, MD	49.7	0.41	0.15	0.081	121.2	1.85	3.1	29.3
6	Salt Creek, NE	32	0.5	0.24	0.038	64	6.32	3.8	52.2
7	Diffcult Run, VA	14.5	0.31	0.25	0.062	46.8	4.03	1.1	1.9
8	Little Pincy Creek, MD	15.9	0.22	0.39	0.053	72.3	7.36	1.4	7.1
9	Bayou Anacoco, LA	17.5	0.45	0.32	0.024	38.9	13.33	2.5	5.8
10	Bayou Anacoco, LA	25.9	0.94	0.34	0.067	27.6	5.07	8.3	32.5
11	Bayou Anacoco, LA	36.6	0.91	0.4	0.067	40.2	5.97	13.3	39.5
12	Comite River, LA	15.7	0.23	0.36	0.039	68.3	9.23	1.3	69
13	Bayou Barthol. LA	33.4	1.4	0.2	0.031	23.9	6.45	9.4	54.7
14	Tickfau River, LA	15	0.59	0.27	0.08	25.4	3.38	2.4	10.3
15	Tangipahoa River, LA	31.4	0.81	0.48	0.072	38.8	6.67	12.2	45.1
16	Red River, LA	253.6	1.62	0.61	0.032	156.5	19.06	250.6	143.8
17	Sabina River, LA	116.4	1.65	0.58	0.054	70.5	10.74	111.4	131.3
18	Sabina River, TX	12.2	0.51	0.23	0.03	23.9	7.67	1.4	14.7
19	Sabina River, TX	21.3	0.93	0.36	0.035	22.9	10.29	7.1	24.2
20	Wind/Big. River, WY	85.3	2.38	1.74	0.153	35.8	11.37	353.2	464.6
21	Wind/Big. River, WY	68.6	2.16	1.55	0.168	31.8	9.23	229.7	162.6
22	Copper Creek, VA	16.7	0.49	0.2	0.08	34.1	2.5	1.6	16.8
23	Clinch River, VA	48.5	1.16	0.21	0.069	41.8	3.04	11.8	14.8
24	Clinch River, VA	57.9	2.45	0.75	0.104	23.6	7.21	106.4	40.5
25	Copper River, VA	19.6	0.84	0.49	0.101	23.3	4.85	8.1	20.8
26	Nooksack River, WA	64	0.76	0.67	0.268	84.2	2.5	32.6	34.8
27	Yadkin River, NC	70.1	2.35	0.43	0.101	29.8	4.26	70.8	111.5
28	Minnesota River	80	2.74	0.034	0.0024	29.2	14.17	7.5	22.3
29	Minnesota River	80	2.74	0.14	0.0097	29.2	14.43	30.7	34.9
30	Amita River	37	0.81	0.29	0.07	45.7	4.14	8.7	23.2
31	Amita River	42	0.8	0.42	0.069	52.5	6.09	14.1	30.2
32	Nooksack River	86	2.93	1.2	0.53	29.4	2.26	302.4	153
33	Susquehanna River	203	1.35	0.39	0.065	150.4	6	106.9	92.9
34	Bayou Anacoco	20	0.42	0.29	0.045	47.6	6.44	2.4	13.9
35	Muddy River	13	0.81	0.37	0.081	16	4.57	3.9	13.9
36	Muddy River	20	1.2	0.45	0.099	16.7	4.55	10.8	32.5

Appendix A: | (*continued*)

No	Stream	B (m)	H (m)	U (m/s)	u_* (m/s)	B/H (−)	U/u_* (−)	Q (m^3/s)	K (m^2/s)
37	Comite River	13	0.26	0.31	0.044	50	7.05	1.0	7
38	Comite River	16	0.43	0.37	0.056	37.2	6.61	2.5	13.9
39	Missouri River	183	2.33	0.89	0.066	78.5	13.48	379.5	465
40	Missouri River	201	3.56	1.28	0.084	56.5	15.24	915.9	837
41	Copper Creek, VA	18.3	0.84	0.52	0.1	21.8	5.2	8.0	21.4
42	Clinc River, TN	46.9	0.86	0.28	0.067	54.5	4.2	11.3	13.9
43	Clinc River, TN	59.4	2.13	0.86	0.104	27.9	8.3	108.8	53.9
44	Copper Crick, VA	18.6	0.39	0.14	0.116	47.7	1.2	1.0	9.9
45	Coachell Canal, CA	24.4	1.56	0.67	0.043	15.6	15.6	25.5	9.6
46	Antietam Creek	15.8	0.39	0.32	0.06	40.5	5.3	2.0	9.3
47	Antietam Creek	24.4	0.71	0.52	0.081	34.4	6.4	9.0	25.6
48	Monocacy River	35.1	0.32	0.21	0.043	109.7	4.9	2.4	4.7
49	Monocacy River	47.5	0.87	0.44	0.07	54.6	6.3	18.2	37.2
50	Elkhorn River	32.6	0.3	0.43	0.046	108.7	9.3	4.2	9.3
51	Sabine River	103.6	2.04	0.56	0.054	50.8	10.4	118.4	315.9
52	Muddy Creek	13.4	0.81	0.37	0.077	16.5	4.8	4.0	13.9
53	Sabine River, TX	35.1	0.98	0.21	0.041	35.8	5.1	7.2	39.5
54	Chattahoochee River	65.5	1.13	0.39	0.075	58.0	5.2	28.9	32.5
Validation set									
55	Antietam Creek, MD	11.9	0.66	0.43	0.085	18	5.06	3.4	20.9
56	Monocacy River, MD	48.7	0.55	0.26	0.052	88.5	5	7.0	37.8
57	Monocacy River, MD	93	0.71	0.16	0.046	131	3.48	10.6	41.4
58	Conococheague Creek, MD	43	1.13	0.63	0.081	38.1	7.78	30.6	53.3
59	Chattahoochee River, GA	75.6	1.95	0.74	0.138	38.8	5.36	109.1	88.9
60	Chattahoochee River, GA	91.9	2.44	0.52	0.094	37.7	5.53	116.6	166.9
61	Bear Creek, CO	13.7	0.85	1.29	0.553	16.1	2.33	15.0	2.9
62	Tangipahoa River, LA	29.9	0.4	0.34	0.02	74.8	17	4.1	44
63	Red River, LA	161.5	3.96	0.29	0.06	40.8	4.83	185.5	130.5
64	Sabina River, LA	160.3	2.32	1.06	0.054	69.1	19.63	394.2	308.9
65	Sabina River, TX	14.2	0.5	0.13	0.037	28.4	3.51	0.9	12.8
66	Wind/Big. River, WY	59.4	1.1	0.88	0.119	54	7.39	57.5	41.8
67	Powell River, TN	33.8	0.85	0.16	0.055	39.8	2.9	4.6	9.5
68	Clinch River, VA	28.7	0.61	0.35	0.069	47	5.07	6.1	10.7
69	Clinch River, VA	53.2	2.41	0.66	0.107	22.1	6.17	84.6	36.9
70	Powell River, TN	36.8	0.87	0.13	0.054	42.3	2.41	4.2	15.5
71	John Day River, OR	25	0.58	1.01	0.14	43.1	7.21	14.6	13.9
72	John Day River, OR	34.1	2.47	0.82	0.18	13.8	4.56	69.1	65
73	Yadkin River, NC	71.6	3.84	0.76	0.128	18.6	5.94	209.0	260.1
74	White River	67	0.59	0.35	0.044	113.6	7.95	13.8	30.2
75	Missouri River	197	3.11	1.53	0.078	63.3	19.62	937.4	892
76	Clinc River, TN	53.3	2.09	0.79	0.107	25.5	7.4	88.0	46.5

Appendix A: | (*continued*)

No	Stream	*B* (m)	*H* (m)	*U* (m/s)	u_* (m/s)	*B/H* (−)	U/u_* (−)	*Q* (m^3/s)	*K* (m^2/s)
77	Antietam Creek	19.8	0.52	0.43	0.069	38.1	6.2	4.4	16.3
78	Monocacy River	36.6	0.45	0.32	0.051	81.3	6.3	5.3	13.9
79	Elkhorn River	50.9	0.42	0.46	0.046	121.2	10.0	9.8	20.9
80	Muddy Creek	19.5	1.2	0.45	0.093	16.3	4.8	10.5	32.5
	Mean	54.3	1.20	0.50	0.084	50.0	7.0	72.7	80.6
	Standard deviation	46.8	1.09	0.37	0.099	31.9	4.57	190.9	176.6
	$X_{\min}$	11.9	0.23	0.034	0.0024	13.8	1.2	0.9	1.9
	$X_{\max}$	253.6	3.96	1.74	0.268	156.5	19.62	892.0	937.4