Predicting Longitudinal Dispersion Coefficient in Natural Streams Using M5# Model Tree

Author

Etemad-Shahidi, Amir, Taghipour, Milad

Published

2012

Journal Title

Journal of Hydraulic Engineering

DOI

https://doi.org/10.1061/(ASCE)HY.1943-7900.0000550

Predicting Longitudinal Dispersion Coefficient in Natural Streams Using M5´ Model Tree

Amir Etemad-Shahidi[1]* and Milad Taghipour[2]

[1*]PhD, School of Civil Engineering, Iran University of Science and Technology,

Narmak, Tehran, Iran, E-mail: etemad@iust.ac.ir, P.O. Box 16765-163, Fax: +9821 77240398.

(Corresponding Author)

M. Sc., School of Civil Engineering, Iran University of Science and Technology,

Narmak, Tehran, Iran, E-mail: miladtaghipour@civileng.iust.ac.ir

ABSTRACT

Longitudinal dispersion coefficient is a key parameter in determining the distribution of pollution concentration; especially in temporally time varying source cases after that full cross sectional mixing has occurred. Several studies have been carried out to present simple formulas for its prediction. However, they may not always result in accurate prediction due to the complexity of the phenomena. In this study, M5´ model tree was used to develop a new model for prediction of the longitudinal dispersion coefficient. The main advantages of the model trees are that they (a) provide transparent formulas and offer more insight into the obtained formulas and (b) are more convenient to develop and employ compared to other soft computing methods. To develop the model tree, extensive field data sets consisting of hydraulic and geometrical characteristics of different rivers were used. The performance of the model was also compared with those of other existing equations using error measures. Overall, results showed that the developed model outperforms the existing formulas and can serve as a valuable tool for prediction of the longitudinal dispersion coefficient.

Key words: longitudinal dispersion; M5´ model tree; spill modeling, river, sinuosity

INTRODUCTION

In rivers, longitudinal dispersion becomes the predominant mechanism in mixing of the tracer by several orders of magnitude when cross sectional mixing is complete; leading to the elimination of any further concentration gradient (Fischer et al. 1979). The dispersion coefficient plays an important role in the spill modeling, design of water intakes, outfall and treatment plants and is representative of the intensity of the mixing in rivers (Deng et al. 2002). Hence, accurate estimation of the longitudinal dispersion coefficient is of a great importance for both engineers and scientists. Direct estimation (by experimental means) of the dispersion coefficient needs expensive and time consuming tracer studies. As a result, demand for a coefficient prediction tool still exists. Estimation of the longitudinal dispersion coefficient has been received considerable attention for a long period of time (e.g. Fischer et al. 1979; Liu 1977; Seo and Cheong 1998; Guymer 1998; Kashefipour and Falconer 2002; Shucksmith et al. 2010). It is still a challenging task to quantify this coefficient since various governing parameters cause complexity in the mixing process. Consequently, introducing mathematical expressions for the dispersion coefficient becomes problematic. Considering that river reaches may vary in condition; one formula may not produce accurate dispersion coefficients. However, this approach is a quite common practice in hydraulic engineering (Rowiński et al. 2005).

When a tracer is introduced to a channel, the shape of tracer cloud is largely affected by velocity variations across the channel. Taylor (1954) suggested that the transverse shear velocity and transverse mixing become in equilibrium after a certain timescale at some point downstream. Beyond this point, Fickian diffusion equation can be used to model the tracer cloud concentration. The following simplified 1-D advection-dispersion equation was derived using Fickian's law for a uniform channel:

$$\left(\frac{\partial C}{\partial t}\right)+U\left(\frac{\partial C}{\partial x}\right)=K_x\left(\frac{\partial^2 C}{\partial x^2}\right) \tag{1}$$

where C is the cross-sectional average concentration (kg/m^3), U is the cross-sectional average velocity (m/s), x is the direction of the mean flow, t is the time (s), and K_x is the longitudinal dispersion coefficient (m^2/s). There is no guarantee for the equilibrium to be established in natural streams. However, equation (1) can adequately illustrate important features of tracer profiles in laboratory and river channels (Rutherford 1994).

Various experimental studies have explored different aspects of the longitudinal dispersion (e.g. Fukuoka and Sayre 1973; Guymer 1998; Murphy et al. 2007). Moreover, regression and dimensional based analysis along with data-driven methods have been employed for the prediction of the dispersion coefficients which have a wide range of variations (e.g. Seo and Cheong 1998; Kashefipour and Falconer 2002; Sahay 2011). More details are provided in the following section.

The main purpose of this study is to employ M5´ algorithm (Wang and Witten 1997) to develop a transparent model for prediction of the longitudinal dispersion coefficient. M5´ model tree is a new soft computing method that provides understandable formulas, which allow users to have more insight in the physics of the phenomena (Etemad-Shahidi and Bonakdar 2009). Rainfall-runoff modeling (Solomatine 2003), flood forecasting (Solomatine and Xue 2004), sediment transport (Bhattacharya and Solomatine 2005) and wave prediction (Etemad-Shahidi and Mahjoobi 2009), are examples of successful model tree applications. This method has not been used for prediction of the dispersion coefficient. In this paper, a comprehensive field data set consisting of 149 field measurements extracted from the literature is used for model development. The performance of the developed model is then compared with those of previous ones using statistical error measures.

PREVIOUS WORKS

In rivers, a range of variables affect the longitudinal dispersion coefficient. The most important ones are: the density, viscosity, channel width, flow depth, mean velocity, shear velocity, bed slope, bed roughness, horizontal stream curvature (sinuosity) and bed shape factor (Seo and Cheong 1998; Guymer 1998). Most of the previous efforts have been devoted to develop a formula for the estimation of K_x using easily measurable parameters such as mean velocity and depth. An overview of these investigations is given first and then a brief report of other affecting parameters (such as sinuosity, vegetation, etc.) and soft computing methods used for the prediction of K_x is provided.

Elder (1959) expanded Taylor's method for an open channel of infinite width. Using laboratory measurements and assuming a logarithmic distribution for the velocity profile in the vertical direction, he suggested:

$$K_x = 5.93 H U_* \tag{2}$$

where H is the depth of flow and U_* is the bed shear velocity. The transverse variation in the velocity profile was not considered in deriving equation (2). This may lead to underestimated predictions since in most natural channels, the transverse shear is more important than the vertical one.

Fischer (1967) used the lateral velocity profile instead of the vertical velocity profile and developed the following integral equation:

$$K_x = -\frac{1}{A}\int_0^W hu'\int_0^y \frac{1}{\varepsilon_t h}\int_0^y hu'dydydy \qquad (3)$$

in which A is the cross-sectional area, W is the channel width, h=$h(y)$ is the local flow depth, u' is the deviation of the velocity from the cross-sectional mean velocity and ε_t is the transverse turbulent diffusion coefficient. This equation shows that K_x is inversely related to ε_t. In narrow and deep rivers, ε_t is high and hence, K_x becomes low. By contrast in relatively wide rivers, the transverse variation of velocity is large and K_x will be higher (Rutherford 1994).

Having difficulties in using the integral form and unavailability of detailed transverse velocity profile, Fischer (1975) simplified equation (3) into the following non-integral form:

$$K_x = 0.011\left(\frac{W^2}{H}\right)\left(\frac{U^2}{U_*}\right) \qquad (4)$$

Liu (1977) (equation 5), Iwasa and Aya (1991) (equation 6) and Koussis and Rodrigues-Mirasol (1998) (equation 7) have considered the effect of the lateral velocity gradient on dispersion and also Fischer's (1975) expression, using laboratory and field data. Their formulas were:

$$\frac{K_x}{HU_*} = \beta\left(\frac{W}{H}\right)^2\left(\frac{U}{U_*}\right)^2, \beta = 0.18\left(\frac{U_*}{U}\right)^{1.5} \qquad (5)$$

$$\frac{K_x}{HU_*} = 2\left(\frac{W}{H}\right)^2 \qquad (6)$$

$$\frac{K_x}{HU_*} = \phi\left(\frac{W}{H}\right)^2, \phi = 0.6 \qquad (7)$$

Koussis and Rodrigues-Mirasol (1998) compared their model with Fischer's (1975) one and stated that their results were much closer to the measurements.

Seo and Cheong (1998) used 59 data sets from rivers in USA. They implemented dimensional analysis to select appropriate variables for model construction and applied one-step Hubor method, a nonlinear multi regression method, to obtain the following equation:

$$\frac{K_x}{HU_*} = 5.915\left(\frac{W}{H}\right)^{0.62}\left(\frac{U}{U_*}\right)^{1.428} \tag{8}$$

They stated that Liu's equation (1977) is generally in good agreement with the measured data whereas Iwasa and Aya's equation (1991) underestimates K_x in many cases.

Deng et al. (2001) developed a mathematical expression for the terms *h*, u' and ε_t from equation (3) and predicted the dispersion coefficient as:

$$\frac{K_x}{HU_*} = 5.915\frac{0.15}{8\varepsilon_{t0}}\left(\frac{W}{H}\right)^{5/3}\left(\frac{U}{U_*}\right)^{2} \text{ for } \frac{W}{H} > 10, \varepsilon_{t0} = 0.145 + \frac{1}{3520}\left(\frac{U}{U_*}\right)\left(\frac{W}{H}\right)^{1.38} \tag{9}$$

Where ε_{t0} is the dimensionless transverse mixing coefficient. Although their model is limited to straight-uniform streams with *W/H* greater than 10, they showed that it is superior to the model of Seo and Cheong (1998) in predicting the K_x. However, the model of Deng et al. (2001) has a disadvantage of the complexity caused by the approximation methods for triple numerical integration with a set of regression equations (Rowiński et al. 2005).

Using 81 sets of field data in USA, Kashefipour and Falconer (2002) developed an equation based on the dimensional and regression analysis as:

$$K_x = 10.612 HU\left(\frac{U}{U_*}\right) \tag{10}$$

They also found out that the average computed ratio of (K_x/HU_*) obtained from Seo and Cheong's (1998) formula and theirs were 1508 and 887 respectively while the corresponding average measured ratios was 1045. Hence, they combined equations (8) and (10) to obtain a more accurate model using trial and error. Their final equation was:

$$K_x = \left[7.428 + 1.775\left(\frac{W}{H}\right)^{0.62}\left(\frac{U_*}{U}\right)^{0.572}\right] HU\left(\frac{U}{U_*}\right) \tag{11}$$

According to their analysis, model of Fischer (1975) and Koussis and Rodrigues-Mirasol (1998) both overestimate the longitudinal dispersion coefficient. They proposed that for open channel flows with *W/H* greater and less than 50, equation (10) and (11) can be used for practical applications, respectively.

In a more fundamental study, Papadimitrakis and Orphanos (2004) stated that the dispersion processes depend on both transverse and vertical velocity profiles, and their relative importance depends on the *W/H* ratio. They divided *W/H* values into three regions and studied each region individually. Various combinations of parameters derived from river geometry and velocity data were tested, and an empirical expression was proposed for different ranges of the *W/H* ratios.

Seo and Baek (2004) developed a theoretical method to predict longitudinal dispersion coefficient based on the distributions of transverse velocity profile in natural streams. First, they tested different velocity profile equations for irregular cross sections. Then, they developed a new equation for the longitudinal dispersion coefficient on the basis of the velocity profile. The comparison showed that the predictions of the developed equation have better agreement with the observed values.

Sahay and Dutta (2009) applied genetic algorithm (GA) to 65 field measurements and proposed:

$$\frac{K_x}{HU_*} = 2\left(\frac{W}{H}\right)^{0.96}\left(\frac{U}{U_*}\right)^{1.25} \tag{12}$$

They mentioned that expressions given by Seo and Cheong (1998), Deng et al. (2001) and Kashefipour and Falconer (2002) perform well especially when the K_x values greater than 100 m^2s^{-1} are excluded from the analysis. They also found that the most effective parameter for an accurate prediction of the longitudinal dispersion coefficient is the term U/U_*.

Tayfur (2009) also used GA approach based on 85 field data and proposed the following empirical equation:

$$K_x = 0.91Q + 9.94 \tag{13}$$

in which Q is the flow discharge. According to this study, equation (13) may have limited predictive capacity for fast-flowing mountainous streams or the streams with a very low flow discharge rate.

Along with these studies, there are some investigations focusing on other influential parameters. For instance, Fukouka and Syre (1973) experimentally investigated the effect of sinuosity in a laboratory flume with various bending conditions. They found that in these cases, the dispersion coefficient is larger and the initial convective period is shorter than those of equivalent straight channel. The effect of this parameter was also investigated by other researchers (e.g. Rutherford 1994; Guymer 1998, Boxall et al. 2003; Boxall and Guymer 2007;

Bashitialshaaer et al. 2011). Effects of other factors such as vegetation, dead zones and hydraulic structures have been studied as well: Nepf et al. (1997) found that the longitudinal dispersion coefficient was decreased in the presence of vegetation while Shucksmith et al. (2010) noticed an increase in the longitudinal mixing in submerged conditions. Valentine and Wood (1977) conducted a numerical modeling to study two-dimensional flow with regular dead zones. They observed that dead zones not only increase the rate of dispersion but also delay the occurrence of Fickian type dispersion. Considerable research efforts have been devoted to the modeling of dead/storage zones in the last decade. More details in this regard can be found in Seo and Cheong (2001), Singh (2003), Smith et al. (2006), Cheong et al. (2007) and Marion et al. (2008). Caplow et al. (2004) suggested that dams (as one of the hydraulic structures) reduce the longitudinal dispersion coefficient below the expected value in a natural channel with the same discharge. However, quantification of the effects of such parameters needs detailed information of the river hydraulics as well as experimental investigations.

Soft computing methods have been also applied by several investigators for the estimation of K_x. Fuzzy logic (Tayfur 2006; Toprak and Savci 2007), adaptive neuro-fuzzy inference system techniques (Riahi-Madvar et al. 2009; Noori et al. 2009), support vector machine (Noori et al. 2009; Azamathulla and Ghani 2010) and genetic programming (Azamathulla and Wu 2011) are the examples of these approaches. It is worth mentioning that artificial neural network (ANN) models have been also employed to predict K_x (Rowiński et al. 2005; Tayfur and Singh 2005; Toprak and Cigizoglu 2008; Sahay 2011).

MATERIAL AND METHOD

Data set

The data sets used in this study were the collection of different data sets measured in different rivers (Fischer 1968; Yotsukura et al. 1970; McQuivey and Keffer 1974; Nordin and Sabol 1974; Rutherford 1994; Graf 1995). By considering the published data sets, 149 distinctive data records were selected which are presented in Appendix A. The data sets contain geometric and hydraulic characteristics including channel width, channel depth, average velocity, shear velocity and longitudinal dispersion coefficient. The histograms of K_x, W/H and U/U_* are illustrated in Fig. 1. Approximately, 80 % of K_x values are less than 100 m^2s^{-1}, the expected maximum value of K_x in natural rivers (Chapra 1997). The histogram of W/H implies that the studied cases varied from narrow rivers (W/H<10) to very wide rivers (W/H>100). U/U_*, defined as the friction term (Seo and Cheong 1998), can be interpreted as hydrodynamic characteristics of the river bed. In other words, the wide range of U/U_* in Fig. 1 covers different bed roughnesses. It should be mentioned that reported coefficients and hydraulic characteristics such as water depth, width and shear velocity may have some uncertainties in their values. Poor estimation procedures, tracer loss or the measurements made in the advective zone are the examples of such uncertainties of K_x values. Besides, software and hardware errors are inevitable in measuring hydraulic characteristics of a river (Rutherford 1994).

Model Tree

The main concept of model tree approach is the process of dividing complex problems into smaller ones (Bhattacharya et al. 2007). Hence, Model tree (MT) can be regarded as a robust method for classification and prediction, which is more understandable than ANN (Jung et al. 2010). In fact, model tree (MT) combines the conventional decision tree with linear regression equation at the leaves (Wang and Witten, 1997). M5 algorithm, initially introduced by Quinlan (1992), is one of the most commonly approaches of MTs. Two main processes are considered in the algorithm: building the tree and deriving the knowledge from it. The first process involves dividing the input parameter space into smaller sub-space for which a multiple regression model is assigned. The scheme is like an inverted tree in which the root is on top while the leaves are at the bottom. In the second process, a data record is introduced into the root of the tree. Fig. 2 illustrates splitting the space for building a tree and eliciting knowledge from the structure.

The record finds its way down by passing through the nodes. Nodes in the tree represent testing the particular parameter. This testing process involves comparing the given parameter with a constant value. These nodes are arranged based on the dividing condition of the first process (the process of building the tree). Finally, related prediction of the introduced record is obtained when a leaf is reached, and it is recognized as an output. Indeed, that record is classified on the basis of the class appointed to that leaf.

M5 algorithm was later improved as M5´ algorithm by Wang and Witten (1997). The new version is more robust, produces simpler trees and can deal with enumerated and missing values. Generally, M5´ consists of three steps: building, pruning and smoothing the tree. M5´ is a recursive algorithm that constructs the regression tree by splitting the space using standard deviation reduction (*SDR*) factor:

$$SDR = sd(T) - \sum_i \frac{T_i}{|T|} \times sd(T_i) \qquad (14)$$

in which T is the set of the data points before splitting, T_i is the data point that results from splitting the space and fall into one sub-space according to the chosen splitting parameter and *sd* is the standard deviation (Wang and Witten 1997). Standard deviation is considered as an error measure for the data points which fall into a one sub-space. M5´ model tree tests different splitting points for all input parameters. For each sub-space, standard deviation is calculated and then compared with that of data records before dividing the space into smaller ones. When a value of the input parameters maximizes the expected error reduction, it is selected as the splitting point (node). This process (splitting) is repeated for every sub-space. The splitting process is brought to an end when a standard deviation reduction is less than 5% or a few data points remain in a sub-domain. After being built, tree calculates a linear multiple regression model for each sub-space using the input parameters.

As the tree grows, the accuracy of the model increases uniformly for training set. Consequently, over-fitting may be inevitable while the tree is being built. Hence, pruning plays an important role in this step. Pruning is the process of merging some of the lower sub-trees into one node to avoid generating too accurate and over-fitted trees. In pruning, the prediction of expected error at each node for the test data is used. The average absolute difference between the predicted value and the actual output is calculated for each of the training sets that reach the node. To prevent underestimating the expected error for new data, the output value is multiplied by $(n+ \nu)/(n- \nu)$ where n is the number of training data points that reach to the node and ν is the number of input parameters that represent the output value at that node. The leaf (or sub-space) can be pruned if the predicted error is lower than the expected one (Witten and Frank 2005).

The last step is the regularization process to compensate sharp discontinuities, which may happen between adjacent linear models in the leaves after the tree is being pruned. In this step, models built in each sup-space calculate the predicted value. That value is then modified along the route back to the root of the tree on top (first splitting point) by smoothing it at each node. The predicted value by the leaf model is combined with that of linear one for each node (Quinlan 1992).

MODELING AND RESULT

As discussed previously (See PREVIOUS WORKS), different parameters can affect longitudinal dispersion coefficient. Considering available data in this study, the effects of some parameters such as vegetation, dead zones and hydraulic structures cannot be investigated. However, we assume that the studied cases here can be representatives of average conditions, which may occur in natural environments. So, the following term relates remaining affecting parameters to K_x:

$$K_x = f_1\left(\rho, \mu, W, H, U, U_*, S_f, \sigma, slope, roughness\right) \tag{15}$$

where ρ is the fluid density, μ is the viscosity, S_f is the bed shape factor and σ is the sinuosity. According to Seo and Cheong (1998), bed shape factor and sinuosity represent the vertical and lateral irregularities respectively.

As mentioned earlier, using dimensional analysis, equation (15) can be written in a dimensionless form as follows (Seo and Cheong 1998; Kashefipour and Falconer 2002):

$$\frac{K_x}{HU_*} = f_2\left(\rho\frac{HU}{\mu}, \frac{W}{H}, \frac{U}{U_*}, S_f, \sigma, slope, roughness\right) \tag{16}$$

in which K_x/HU_* is the dimensionless dispersion coefficient and $\rho HU/\mu$ is the Reynolds number. Since the flow in natural rivers is usually turbulent, the effect of Reynolds number is negligible and can be ignored. The effects of channel slope and roughness can be reflected in terms *of* U_* and U/U_* respectively and can be excluded.

Because of the complexity of obtaining σ and also the limited number of available data for this parameter, it has been omitted in most of the previous studies. However, some investigators commented on the effect of σ. Sahay (2011), Tayfur and Singh (2005) and Rowiński et al. (2005) stated that the inclusion of σ in the input vector of the ANN models improves the accuracy of prediction. On the contrary, Tayfur (2006) stated that there is no strong dependence between K_x and σ. In fact, our study has also confirmed the former finding. In this study, σ was reported in about 40% of the whole measurements. Hence, σ was excluded from the input parameters of the model tree at the first step to simplify the problem.

S_f values are not reported in the data sets, and hence it was not possible to use it. This parameter is not easily collected from natural streams, and its corresponding effect can be included in the term U/U_* (Seo and Cheong 1998). However, Deng et al (2001) introduced an expression named channel shape parameter as $\beta=\ln(W/H)$. β might be able to reflect the vertical irregularities as the bed shape factor. As seen later, one of the inputs of our model is log (W/H), which corresponds to β. Hence equation (16) can be written as:

$$\frac{K_x}{HU_*} = f_3\left(\frac{W}{H}, \frac{U}{U_*}\right) \tag{17}$$

Assuming f_3 to be a power function, the general expression of longitudinal dispersion coefficient can be:

$$\frac{K_x}{HU_*} = a\left(\frac{W}{H}\right)^b \left(\frac{U}{U_*}\right)^c \tag{18}$$

in which *a, b, c* are the constants of the equation which possess different values in different expressions.

Since, model trees ordinarily can only produce linear relationships, the model was developed with log (*inputs)* and log (*output*) to obtain a nonlinear relationship. Furthermore, most of the data-driven approaches perform well while dealing with the data with nearly uniform or normal distributions (Pyle 1992). It can be easily inferred from Fig. 3 that the distributions of the used variables are nearly log normal.

Considering possible combinations of dimensionless forms for the longitudinal dispersion coefficient, plots of (K_x/HU_*), (K_x/HU), (K_x/WU_*), (K_x/WU) versus W/H and U/U_* were plotted and their correlation coefficients were calculated. It was found that (K_x/HU_*) is the best dimensionless form of K_x and has the highest correlation with W/H and U/U_*. Consequently, these three terms were used as the inputs and the output for the model development.

Taking logarithms of equation (18), the following linear formula can be derived:

$$\log(\frac{K_x}{HU_*}) = \log a + b\log\left(\frac{W}{H}\right) + c\log\left(\frac{U}{U_*}\right) \qquad (19)$$

To develop the model, test and train technique was used. This is a common technique in learning algorithms on a data set (Mahjoobi et al. 2008). In this method, a data set is randomly divided into two subsets, train and test. The train data set is used to train the model, and then the model is tested (verified) using the test data set. In this study, 119 data records were used for training while the remaining ones were used for testing the model. The statistics of the parameters used for training the model are listed in Table 2. The developed model tree (MT) generated the following formulas:

If log (W/H) $\leq$ 1.486 then log (K_x/HU_*) = 1.90+0.78log (W/H) + 0.11log (U/U_*) 20-a

If log (W/H) > 1.486 then log (K_x/HU_*) = 1.15+0.61log (W/H) + 0.85log (U/U_*) 20-b

After transformation, Equations (20-a, 20-b) can be written as:

If $W/H \leq 30.6$ then $\left(\frac{K_x}{HU_*}\right) = 15.49\left(\frac{W}{H}\right)^{0.78}\left(\frac{U}{U_*}\right)^{0.11}$ 21-a

If $W/H > 30.6$ then $\left(\frac{K_x}{HU_*}\right) = 14.12\left(\frac{W}{H}\right)^{0.61}\left(\frac{U}{U_*}\right)^{0.85}$ 21-b

The splitting parameter is *W/H*, and the splitting value is about 30; a value close to that obtained by Papadimitrakis and Orphanos (2004). This splitting value is obtained by minimizing the prediction error and do not necessarily have a physical interpretation (Bhattacharya et al. 2007; Bonakdar and Etemad-Shahidi 2011). However, the importance of *W/H* in determining K_x has been mentioned by others (Asay and Fujisaki 1991; Kashefipour and Falconer 2002; Papadimitrakis and Orphanos 2004, Tayfur and Singh 2005). Transverse shear is less important in relatively small values of *W/H*, while it dominates the dispersion characteristic when the aspect ratio is large. Hence, different regimes may exist for low and high *W/H* ratios.

The exponents of *W/H* and U/U_* are different in these formulas. In rivers with $W/H \leq 30.6$, it is the width to depth ratio that outweighs the dispersion coefficient. In wider rivers with $W/H > 30.6$, the influence of U/U_* increases and the effect of *W/H* decreases (see also Papadimitrakis and Orphanos 2004). This can be interpreted as in very wide rivers, K_x may be less influenced by the *W/H* ratio than in narrow rivers. As discussed by Rutherford (1994), in relatively wide rivers, the role of velocity becomes more pronounced in determining the K_x than in narrow rivers. In this regard, it can be seen that the power of *U* in equation 21-b is almost 8 times greater than that of the value in equation 21-a. The obtained exponents of *W/H* and U/U_* are in the range reported in previous works. The average exponents of *W/H* and U/U_* are 0.7 and 0.48, which are close to those of Seo and Cheong's (1998) and Liu's (1977), respectively. In brief, it was concluded that the obtained formulas are in good agreement with the engineering sense and previous findings.

The performance of the developed model was evaluated against those of other existing models by using error measures such as discrepancy ratio (*DR*) (White et al. 1973), mean of absolute error (*ME*) and root mean square (*RMS*). These parameters are defined as:

$$DR = \log \frac{K_{x_p}}{K_{x_m}} \tag{22}$$

$$ME = \frac{1}{N} \sum_{i=1}^{N} \left| DR_i \right| \tag{23}$$

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(DR_i \right)^2} \tag{24}$$

in which K_{x_p} and K_{x_m} are predicted and measured dispersion coefficients, respectively and *N* is the total number of data points.

If *DR* is equal to zero, there will be an exact matching between the measured and predicted values. Otherwise, there is either an overestimation (*DR*>0) or underestimation (*DR*<0). Accuracy is defined as the percentage of *DR* values, which fall between -0.3 to 0.3 (Seo and Cheong 1998; Kashefipour and Falconer 2002). The performance of each model can also be determined by comparing the calculated values of *ME*, *RMS* with zero. The closer the values to zero, the more accurate the model will be.

Error measures of previous models and the developed one are presented in Table 2. The results in the last two rows of Table 2 show that the errors of the developed model for testing data as well as all data are nearly the same. Although Elder's (1959) equation is more suitable for the rivers with no transverse shear, the comparison of this equation with others can merely illustrate the importance of transverse variation. According to Table 2, the performance of Fischer's model (1975) is the least satisfactory after Elder (1959). As seen, all the error measures of the developed model show improvement in prediction of the longitudinal dispersion coefficient. The MT has the accuracy of 63%, the highest one among others. The nearest value of the accuracy to that of MT belongs to Liu (1977) with about 51%. The difference between these two accuracy values shows superiority of MT over other models well. *ME* and *RMS* are other performance indicators. The developed model outperforms other ones as it has the lowest values for these two error measures. Besides, the percentage of *DR* values greater than 0.3 and less than -0.3 of MT are 17.4% and 19.5%, respectively. It means that *DR* values out of this range are almost equally distributed between overestimated and underestimated values. But for the other models, non symmetric distributions for values out of the range -0.3 to 0.3 are somehow considerable. Liu (1977), Seo and Cheong (1998), Deng et al. (2001) and Sahay and Dutta (2009) over predict the dispersion coefficient by 1.7 times more than the under predicted cases. In other words, they generally overestimate the measured values of the longitudinal dispersion coefficient. The overestimation of the longitudinal dispersion coefficient results in obtaining lower maximum concentration. This is an important issue, especially in practical application regarding maximum concentration estimation. In such cases, it may not be safe to use over predicted K_x values.

The histograms of *DR* values for six models are compared in Fig. 4. The *DR* distribution of MT shows a nearly symmetrical distribution between -1 to 1, which means there is relatively no skewness towards positive or negative values. However, there are cases of overestimation for other models. For example, model of Sahay and Dutta (2009) is skewed to positive values and does not have a symmetrical distribution. This can also be understood from Table 3 in which the mean, standard deviation and skewness of *DR* of different models are given. The skewness values of our model and Deng et al.'s (2002) are the lowest ones. Moreover, the mean *DR* of MT is zero which is an indication of a symmetrical distribution. About 68 % of the *DR* values of Sahay and Dutta's (2009) fall between -0.41 and 0.63, and 95% of the values are between -0.93 and 1.15, implying skewness towards positive values. As another example, model of Kashefipour and Falconer II (2002) can also be addressed. Although the corresponding mean value of *DR* is close to zero, it has a relatively high standard deviation. As seen, about 95% the *DR* values are between -0.94 and 1.15, whereas that of MT is in the range of -0.86 and 0.86. This implies that *DR* values of MT prediction are closer to zero. Table 4 also shows that, MT has the lowest maximum error while Kashefipour and Falconer II (2002) has the largest one.

In addition, the correlation coefficient (*CC*) and the slope of the regression line are other tools for evaluating the performance of a model. If the slope of the regression line for prediction versus measured data is close to 1 and the value of *CC* is high, then the model is accurate. The developed model greatly outperforms other ones in predicting K_x when extreme measured values of dispersion coefficient (K_x>100 m^2s^{-1}) are excluded from the analysis. As presented in Table 5, the slope of the regression line of MT is close to 1 and it has the highest *CC*.

Some more information was gained by the introduction of model tree equations. Information such as the splitting point and its value and the values of the exponent of the input parameters helped us to include σ parameter for the cases in which it was reported. Considering equations 21-a, 21-b and the splitting point, the data set was divided into two sub sets. For each sub set, the effect of σ was considered to be a power function ($\lambda = A\sigma^B$). The constants were obtained by non-linear regression relation using the reported data (including σ). Finally, equations 21-a and 21-b were modified as:

If $W/H \leq 30.6$ then $$\left(\frac{K_x}{HU_*}\right) = 2.75\left(\frac{W}{H}\right)^{0.78}\left(\frac{U}{U_*}\right)^{0.11}(\sigma)^{4.04} \qquad \text{25-a}$$

If $W/H > 30.6$ then $$\left(\frac{K_x}{HU_*}\right) = 8.36\left(\frac{W}{H}\right)^{0.61}\left(\frac{U}{U_*}\right)^{0.85}(\sigma)^{1.70} \qquad \text{25-b}$$

These equations show that K_x is directly related to the sinuosity which is in line with the previous findings of Fukouka and Syre (1973) and Bashitialshaaer et al. (2011). Interestingly, the performance of the modified equations was improved when including sinuosity. Table 5 summarizes error measures of equations 21 and 25 for both ranges and also for all data with reported σ values. It is found that the accuracy of new equations accounting for σ was enhanced especially for the lower range of *W/H*. It is noteworthy that the power of σ depends on the *W/H* ratios. As inferred from eq. 25, the sinuosity has a greater effect on the lower *W/H* ratios than in the higher ones. This is in good agreement with engineering sense since in narrow rivers, mixing is more influenced by the river curvatures.

Application of the piece-wise regression might provide better understanding of the physics of the phenomena in comparison with one simple equation, which may not be appropriate for all cases. However, comparison of two equations with existing ones is inevitable for the illustration of performance of the new model. Model tree approach used in this study requires minimum effort in comparison with other soft computing methods. The model tree provides simple regression formulas with low computational cost (Jafari and Etemad-Shahidi 2011). By contrast to other soft computing methods such as ANN, it does not need too much trial and error for obtaining the best model. Besides, it is more transparent and can provide understandable formulas. The latter advantage can benefit users to have more insight into the physics of the phenomena along with quantifying the role of each input parameter. Other soft computing such as ANN have limited applicability because they are more like a black box model and do not reveal any direct mathematical expressions (Tayfur 2006). Model trees have some limitations as well. As mentioned before, they can only produce linear relationships. Besides, transformation of input parameters may not be that simple in more complex cases and not necessarily lead to a few simple linear formulas.

CONCLUSION

In this study, M5´ model tree was used to predict the longitudinal dispersion coefficient in natural streams. The model was developed using 149 field data records consisting of hydraulic and geometrical characteristics. Because of the limited number of reported values of σ, it was decided to develop the equations without σ in the first step. Based on previous studies and trial and error, W/H and U/U_* along with (K_x/HU_*) were used as inputs and output of the model tree, respectively. Two formulas were generated and the splitting parameter was W/H, which is an important parameter in dispersion mechanism. The performance of the new model was evaluated, and the results were compared with those of existing formulas using different error measures. The developed model outperformed others in terms of accuracy. Effect of σ was considered then and the results showed improvement in the predictions of the dispersion coefficient. The suggested models seem to be safely applicable in hydraulic and environmental studies such as design of outfalls or evaluating risks from spills of hazardous contaminants.

ACKNOWLEDGEMENTS

The authors acknowledge comments of Professor Jorg Imberger on the earlier version of the manuscript. We also thank Meysam Bali, Ebrahim Jafari, Ali Behnood and Amin Asgarian for their help in improving the manuscript.

NOTATION

The following symbols are used in this paper:

A = cross-sectional area;

C = cross-sectional average concentration;

H = depth of flow;

h = local flow depth;

K_x = longitudinal dispersion coefficient;

K_{x_p} = measured dispersion coefficient;

K_{x_m} = predicted dispersion coefficient;

Q = flow discharge;

n= number of train data points;

N= number of data points;

sd = standard deviation;

S_f = bed shape factor;

t = time;

T = set of the examples that reach the node;

T_i = set of the results of the node splitting according to selected parameter;

U = cross-sectional average velocity;

U_* = bed shear velocity;

u' = deviation of the velocity from the cross-sectional mean velocity;

υ = number of inputs

x = direction of the mean flow;

ρ= fluid density;

μ = fluid viscosity;

σ= sinuosity;

β=channel shape parameter

ε_t = transverse turbulent diffusion coefficient;

ε_{t0} = dimensionless transverse mixing coefficient;

REFERENCES

Asay, K., and Fujisaki, K. (1991). "Effect of aspect ratio on longitudinal dipsersion coefficient." *Proc., Int. Symp. on Envir. Hydr.*, A. A. Balkema, Rotterdam, The Netherlands, 493-498.

Azamathulla, H. M., and Ghani, A. A. (2010). "Genetic Programming for Predicting Longitudinal Dispersion Coefficients in Streams." *Wat. Res. Manag*., 1-8.

Azamathulla, H. M., and Wu, F. C. (2011). "Support vector machine approach for longitudinal dispersion coefficients in natural streams." *Appl. Soft Comp.*, 11(2), 2902-2905.

Bashitialshaaer, R., Bengtsson, L., Larson, M., Persson, K. M., Aljaradin, M., and Hossam I, A. (2011) "Sinuosity effects on Longitudinal Dispersion Coefficient*", Int. J. of Sust Wat Environ Sys.*, 2(2), 77-84

Bhattacharya, B., Price, R. K., and Solomatine D. P. (2007). "Machine learning approach to modeling sediment transport." *J. Hydraul. Eng.,* 133(4), 440-450.

Bhattacharya, B., and Solomatine, D. P. (2005). "Neural networks and M5 model trees in modelling water level-discharge relationship." *Neurocomp.*, 63, 381-396.

Bonakdar, L. and Etemad-Shahidi, A. (2011). "Predicting wave run-up on rubble-mound structures using M5 ′ machine learning method." *Ocean Eng.*, 38(3), 111-118.

Boxall, J. B. and Guymer, I. (2007) "Longitudinal mixing in meandering channels: New experimental data set and verification of a predictive technique", *Wat. Re*s., 41(2), 341-354

Boxall, J. B., Guymer, I. and Marion, A. (2003) "Transverse mixing in sinuous natural open channel flows", *J. Hydraul Res.*, 41(2), 153-165

Caplow, T., Schlosser, P. and Ho, D. T. (2004) "Tracer study of mixing and transport in the upper Hudson River with multiple dams", *J. Environ. Eng.*, 130 (12), 1498-1506

Chapra, S. C. (1977). *Surface Water-Quality Modeling*, McGraw-Hill, 845

Cheong, T. S., Younis, B. A., and Seo, I. W. (2007). “Estimation of key parameters in model for solute transport in rivers and streams.” *Water Resour. Manage*., 21(7), 1165-1186.

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

Deng, Z. Q., Singh, V. P., and Bengtsson, L. (2001). “Longitudinal dispersion coefficient in straight rivers.” *J. Hydraul. Eng.,* 127(11), 919-927.

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

Etemad-Shahidi, A., and Bonakdar, L. (2009). “Design of rubble-mound breakwaters using M5 ′ machine learning method.” *Appl. Ocean Res.*, 31(3), 197-201.

Etemad-Shahidi, A., and Mahjoobi, J. (2009). “Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior.” *Ocean Eng.*, 36(15-16), 1175-1181.

Fischer, B. H. (1967). “The machanics of dispersion in natural streams.” *J. Hyraul. Div., ASCE*, 93(6), 187-216.

Fischer, B. H. (1968). “Dispersion predictions in natural streams.” *J. Saint. Eng. Div., ASCE*, 94(5), 927-943.

Fischer, B. H. (1975). “Discussion of ‘simple method for predicting dispersion in streams.’ by R. S. McQuivey and T. N. Keefer.” *J. Environ. Eng. Div., ASCE,* 101(3), 543-455.

Fischer, B. H., List, E. J., Koh, R. C. Y., Imberger, J., and Brooks, N. H. (1979). *Mixing in land and Costal Waters*, Academic, New York, 104-138.

Fukuoka, S. and Sayre, W. W. (1973) “Longitudinal dispersion in sinuous channels”, *.J. Hyraul. Div., PROC. ASCE,* 99(1), 195-217

Graf, B. (1995). “Observed and predicted velocity and longitudinal dispersion at steady and unsteady flow, Colorado River, Glen Canyon Dam to Lake Mead.” *Water Resour. Bull.,* 31(2), 265-281.

Guymer, I. (1998) “Longitudinal dispersion in sinuous channel with changes in shape”, *J. Hydraul. Eng.,*, 124(1), 33-40

Iwasa, Y., and Aya, S. (1991).”Predicting longitudinal dispersion coefficient in open-channel flows.” *Proc., Int. Symp on Envir. Hydr.,* Hong Kong, 505-510.

Jafari, E., and Etemad-Shahidi, A. (2012). “Derivation of a new model for prediction of wave overtopping at rubble-mound structures”, *J. Waterway, Port and Coastal Eng.,*under press

Jung, N. C., Popescu, I., Kelderman, P., Solomatine, D. P., and Price, R. K. (2010). “Application of model trees and other machine learning techniques for algal growth prediction in yongdam reservoir, Republic of Korea.” *J. Hydroinf.,* 12(3), 262-274.

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

Koussis, A. D., and Rodríguez-Mirasol, J. (1998). “Hydraulic estimation of dispersion coefficient for streams.” *J. Hydraul. Eng.,* 124(3), 317-320.

Liu, H. (1977). “Predicting dispersion coefficient of streams.” *J. Environ. Eng. Div., ASCE,,* 103(1), 59-69.

Mahjoobi, J., Etemad-Shahidi, A., and Kazeminezhad, M. H. (2008). “Hindcasting of wave parameters using different soft computing methods.” *Appl. Ocean Res.,* 30(1), 28-36.

Marion, A., Zaramella, M., and Bottacin-Busolin, A. (2008). “Solute transport in rivers with multiple storage zones: The STIR model.” *Water Resour. Res.*, 44(10), 1-21.

McQuivey, R. S., and Keffer, T. N. (1974). “Simple method for predicting dispersion in streams.” *J. Environ. Eng. Div., ASCE,* 100(4), 997-1011.

Murphy, E., Ghisalberti, M. and Nepf, H. (2007) “Model and laboratory study of dispersion in flows with submerged vegetation”, *Water Resour. Res*., 43(5)

Nepf, H. M., Mugnier, C. G. and Zavistoski, R. A. (1997) “The effects of vegetation on longitudinal dispersion”, *Estuar Coast Shelf S.*, 44:6, 675-684

Noori, R., Karbassi, A. R., Farokhnia, A., and Dehghani, M. (2009). “Predicting the Longitudinal Dispersion Coefficient Using Support Vector Machine and Adaptive Neuro-Fuzzy Inference System Techniques.” *Env. Eng. Sci.*, 26(10), 1503-1510.

Nordin, C. F., and Sabol, G. V. (1974).”Empirical data on longitudinal dispersion in rivers.” *U.S. Geological Survey Water Resour. Investigation 20-74*, Washington, D. C.,

Papadimitrakis, I., and Orphanos, I. (2004). “Longitudinal dispersion characteristics of rivers and natural streams in Greece.” *Wat., Air, and Soil Pol.: Focus.,* 4(4-5), 289-305.

Pyle, D. (1992). *Data preparation for data mining*, Morgan Kaufmann, Calif.,

Quinlan, J. R. (1992).”Learning with continuous classes.” *Proc.,5th Australian Joint Conf. on Artificial Intelligence,* World Scientific, Singapore, 343-348.

Riahi-Madvar, H., Ayyoubzadeh, S. A., Khadangi, E., and Ebadzadeh, M. M. (2009). “An expert system for predicting longitudinal dispersion coefficient in natural streams by using ANFIS.” *Exp. Sys. with Appl.,* 36(4), 8589-8596.

Rowiński, P. M., Piotrowski, A., and Napiórkowski, J. J. (2005). "Are artificial neural network techniques relevant for the estimation of longitudinal dispersion coefficient in rivers?., *Hyrolo. Sci. J.*, 50(1), 175-187.

Rutherford, J. C. (1994). *River Mixing*, Wiley, Chichester, U. K, 347

Sahay, R. R. (2011). "Prediction of longitudinal dispersion coefficients in natural rivers using artificial neural network." *J. Fluid Mech*., 11(3), 247-261

Sahay, R. R., and Dutta, S. (2009). "Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm." *Hydrol. Res.*, 40(6), 544-552.

Seo, I. W. and Baek, K. O. (2004). "Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams" *J. Hydraul. Eng.*, 130(3), 227-236.

Seo, I. W., and Cheong, T. S. (2001). "Moment-based calculation of parameters for the storage zone model for river dispersion. " *J. Hydraul. Eng.*, 127(6), 453-465.

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

Shucksmith, J. D., Boxall, J. B. and Guymer, I. (2010) "Effects of emergent and submerged natural vegetation on longitudinal mixing in open channel flow". *Water Resour. Res.*, 46(4), 1-14.

Singh, S. K. (2003). "Treatment of stagnant zones in riverine advection-dispersion. " *J. Hydraul. Eng.*, 129(6), 470-473.

Solomatine, D. P. (2003). "Model trees as an alternative to neural networks in rainfall-runoff modelling." *Hyrolo. Sci. J.,* 48(3), 399-411.

Solomatine, D. P., and Xue, Y. P. (2004). “M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China.” *J. Hydrol. Eng.*, 9(6), 491-501.

Smith, P., Beven, K., Tawn, J., Blazkova, S., and Merta, L. (2006). “Discharge-dependent pollutant dispersion in rivers: Estimation of aggregated dead zone parameters with surrogate data. ” *Water Resour. Res.*, 42(4), 1-9.

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

Tayfur, G. (2009). “GA-optimized model predicts dispersion coefficient in natural channels.” *Hydrol. Res.,* 40(1), 65-75.

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

Taylor, G. I. (1954).”The dispersion of matter in turbulent flow through a pipe.” *Proc. R. Soc., London Ser. A*, 223, 446-468.

Toprak, Z. F., and Cigizoglu, H. K. (2008). “Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods.” *Hydrol. Proc.*, 22(20), 4106-4129.

Toprak, Z. F., and Savci, M. E. (2007). “Longitudinal dispersion coefficient modeling in natural channels using fuzzy logic” *Clean-Soil Air Wat.*, 35(6), 626-637.

Valentine, E. M., and Wood, I. R. (1977). “Longitudinal dispersion with dead zones”, *J. Hyraul. Div., ASCE.*, 103(9), 975-990.

Wang, Y. and Witten, I. H. (1997).”Induction of model trees for predicting continuous lasses”. *Proc of the Poster Papers of the European Conference on Machine Learning, 1997* Prague. University of Economics, Faculty of Informatics and Statistics.

White, W. R., Milli, H., and Crabbe, A. D. (1973).”Sediment transport an appraisal methods, Vol 2: Performance of theoretical methods when applied to flume and field data.” *Hyr. Res. Station Rep., No. 1T119*, Wallingford, U. K.,

Witten, I. H. and Frank, E. (2005). ”Data Mining- Practical Machine Learning Tools and Techniques”, Morgan kaufmann, San Francisco.

Yotsukura, N., Fischer, H. B., and Sayre, W. W. (1970).”Measurement of mixing characteristics of the Missouri River between Sioux City, Iowa and Plattsmouth, Nebraska.” *U.S. Geological Survey Water-Supply paper 1899-G*, Washington, D. C.,

Table 1 The statistics of parameters used for training the model

	W (m)	H (m)	U (ms^{-1})	U_* (ms^{-1})	W/H	U/U_*	K_x (m^2s^{-1})	K_x/U_*H
Max	253.6	8.2	1.73	0.55	403.75	20.25	1486.5	37140
Min	1.4	0.14	0.03	0.002	2.20	0.77	0.2	3.08
Avg	48.6	1.36	0.48	0.087	47.72	6.96	79.4	1172
Std	47.2	1.39	0.33	0.078	49.64	4.75	174.9	3570

Table 2 Comparison of various models performance

Model	*DR*<-0.3	-0.3< *DR* <0	0< *DR* <0.3	*DR* >0.3	Accuracy	*ME*	*RMS*
Elder (1959)-all data	98.0	1.3	0.7	0.0	2.0	1.85	1.95
Fischer (1975) -all data	30.2	18.1	16.8	34.9	34.9	0.56	0.71
Liu (1977),all data	17.4	22.1	28.9	31.6	51.0	0.42	0.57
Seo and Cheong (1998), all data	18.8	16.1	30.2	34.9	46.3	0.43	0.59
Deng et al (2001), all data	20.1	19.5	27.5	32.9	47.0	0.42	0.56
Kashefipour and Falconer (2002), all data	36.9	30.2	10.7	22.2	40.9	0.54	0.74
Kashefipour and Falconer (2002)-II, all data	26.1	29.5	19.4	25	48.9	0.46	0.66
Sahay and Dutta (2009), all data	20.1	22.8	22.8	34.3	45.6	0.40	0.53
MT, all data	17.4	28.9	34.2	19.5	63.1	0.32	0.44
MT, Testing data	13.3	36.7	26.7	23.3	63.3	0.35	0.48

Table 3 Mean ($\overline{DR}$), standard deviation (σ_{DR}), maximum ($\left|\overline{DR_{\max}}\right|$) and skewness of *DR* ($SK_{DR}$) of different models

Model	$\overline{DR}$	σ_{DR}	$\left|\overline{DR_{\max}}\right|$	SK_{DR}
Liu(1977)	0.11	0.56	2.05	-0.52
Seo and Cheong (1998)	0.18	0.56	2.15	0.8
Deng et al. (2001)	0.11	0.56	1.89	0.33
Kashefipour and Falconer (2002)II	0.03	0.66	2.74	1.1
Sahay and Dutta (2009)	0.11	0.52	1.79	0.38
MT	0.00	0.43	1.54	-0.32

Table 4 Slope of regression line and *CC* for prediction versus measured K_x (K_x<100 m^2s^{-1}) of different models

Model	Slope of regression line	*CC*
Liu (1977)	3.86	0.29
Seo and Cheong (1998)	1.31	0.46
Deng et al. (2001)	0.61	0.14
Kashefipour and Falconer II (2002)	0.51	0.18
Sahay and Dutta(2009)	1.4	0.49
MT	0.96	0.6

Table 5 Comparison of error measures of MT equations with and without σ

	W/H < 30.6		*W/H* > 30.6		All data points with reported σ value	
	Without	With	Without	With	Without	With
MAE	0.53	0.33	0.25	0.2	-0.07	0
RMSE	0.71	0.42	0.31	0.27	0.44	0.32
$\overline{DR}$	-0.23	0	0	0	0.32	0.24
σ_{DR}	0.7	0.44	0.32	0.27	0.44	0.31

Appendix A. The data sets used in this study.

No	Stream	$W(m)$	$H(m)$	$U(m/s)$	$U_*(m/s)$	$K_x\ (m^2/s)$	σ
1	Copper creek, VA(below gage)	15.9	0.49	0.21	0.079	19.52	
2	Copper creek, VA(below gage)	18.3	0.84	0.52	0.1	21.4	
3	Copper creek, VA(below gage)	16.2	0.49	0.25	0.079	9.5	
4	Clinch river, TN(below gage)	46.9	0.86	0.28	0.067	13.93	
5	Clinch river, TN(below gage)	59.4	2.13	0.86	0.104	53.88	
6	Clinch river, TN(below gage)	53.3	2.09	0.79	0.107	46.45	
7	Copper creek, VA(above gage)	18.6	0.39	0.14	0.116	9.85	
8	Power river, TN	33.8	0.85	0.16	0.055	9.5	
9	Clinch river, VA	36	0.58	0.3	0.049	8.08	
10	Green and Duwamish	21.77	1.58	0.31	0.058395	6.5	
11	Green and Duwamish	29.61	1.08	0.36	0.048279	0.5	1.41
12	Bayou Anacoco	19.8	0.41	0.29	0.044	13.94	1.30
13	Nooksack river	86	2.94	1.2	0.514	153.29	
14	Antietam creek	15.8	0.39	0.32	0.06	9.29	
15	Antietam creek	19.8	0.52	0.43	0.069	16.26	
16	Antietam creek	24.4	0.71	0.52	0.081	25.55	
17	Monocacy river	35.1	0.32	0.21	0.04	4.65	
18	Monocacy river	36.6	0.45	0.32	0.05	13.94	
19	Monocacy river	47.5	0.87	0.44	0.07	37.16	
20	Missouri river	182.9	2.23	0.93	0.065	464.52	1.35
21	Missouri river	201.2	3.56	1.27	0.082	836.13	1.35
22	Missouri river	196.6	3.11	1.53	0.077	891.87	1.35
23	Wind/Bighom rivers	67.1	0.98	0.88	0.11	41.81	
24	Elkhom river	32.6	0.3	0.43	0.046	9.29	
25	Elkhom river	50.9	0.42	0.46	0.046	20.9	
26	John day river	25	0.56	1.01	0.137	13.94	1.08
27	Comite river	12.5	0.26	0.31	0.043	6.97	1.31
28	Comite river	15.8	0.41	0.37	0.055	13.94	1.31
29	Amite river	36.6	0.81	0.29	0.068	23.23	
30	Amite river	42.4	0.8	0.42	0.068	30.19	
31	Sabine river	103.6	2.04	0.56	0.054	315.87	
32	Sabine river	127.4	4.75	0.64	0.081	668.9	
33	Muddy creek	13.4	0.81	0.37	0.077	13.94	
34	Muddy creek	19.5	1.2	0.45	0.093	32.52	
35	Sabine river Texas	35.1	0.98	0.21	0.041	39.48	
36	white river	67.1	0.55	0.35	0.044	30.19	
37	Chattahoochee river	65.5	1.13	0.39	0.075	32.52	
38	Susquehanna river	202.7	1.35	0.39	0.065	92.9	1.13

39	Antietam Creek	10.97	0.52	0.21	0.074909	17.5	
40	Antietam Creek	23.47	0.7	0.52	0.101491	101.5	
41	Antietam Creek	24.99	0.45	0.41	0.081374	25.9	
42	Antietam Creek	12.8	0.3	0.42	0.057	17.5	1.40
43	Antietam Creek	24.08	0.98	0.59	0.098	101.5	2.25
44	Antietam Creek	11.89	0.66	0.43	0.085	20.9	2.25
45	Antietam Creek	21.03	0.48	0.52	0.069	25.9	1.26
46	Monocacy river	48.7	0.55	0.26	0.05	37.8	1.28
47	Monocacy river	92.96	0.71	0.16	0.05	41.4	1.28
48	Monocacy river	51.21	0.65	0.62	0.04	29.6	1.28
49	Monocacy river	97.54	1.15	0.32	0.058	119.8	1.61
50	Monocacy river	49.99	0.95	0.32	0.074778	29.6	
51	Monocacy river	33.53	0.58	0.16	0.041315	66.5	
52	Monocacy river	40.54	0.41	0.23	0.04	66.5	1.61
53	Conococheague Creek	42.21	0.69	0.23	0.064	40.8	2.25
54	Conococheague Creek	49.68	0.41	0.15	0.081	29.3	2.25
55	Conococheague Creek	42.98	1.13	0.63	0.081	53.3	1.31
56	Conococheague Creek	43.28	0.69	0.22	0.063729	40.8	
57	Conococheague Creek	63.7	0.46	0.1	0.056203	29.3	
58	Conococheague Creek	59.44	0.76	0.68	0.072242	53.3	
59	Chattahoochee river	75.6	1.95	0.74	0.138	88.9	1.27
60	Chattahoochee river	91.9	2.44	0.52	0.094	166.9	1.57
61	Chattahoochee river	99.97	2.5	0.3	0.105054	166.9	
62	Salt Greek	32	0.5	0.24	0.038	52.2	1.38
63	Difficult run	14.5	0.31	0.25	0.062	1.9	1.09
64	Difficult run	11.58	0.4	0.22	0.087475	1.9	
65	Bear Creek	13.7	0.85	1.29	0.553	2.9	1.08
66	Little Pincy Creek	15.9	0.2	0.39	0.053	7.1	1.13
67	Bayou Anacoco	17.5	0.45	0.32	0.024	5.8	1.41
68	Bayou Anacoco	25.9	0.94	0.34	0.067	27.6	1.41
69	Bayou Anacoco	36.6	0.91	0.4	0.067	40.2	1.41
70	Comite river	15.7	0.2	0.36	0.04	69	1.31
71	Comite river	6.1	0.49	0.25	0.057591	69	
72	Bayou Bartholomew	33.4	1.4	0.2	0.03	54.7	2.46
73	Bayou Bartholomew	37.49	2.07	0.1	0.040306	54.7	
74	Amite river	21.3	0.5	0.54	0.027	501.4	
75	Amite river	46.02	0.53	0.41	0.042659	501.4	
76	Tickfau River	14.9	0.59	0.27	0.08	10.3	1.75
77	Tickfau River	41.45	1.04	0.07	0.090343	10.3	
78	Tangipahoa River	31.4	0.81	0.48	0.072	45.1	1.46
79	Tangipahoa River	29.9	0.4	0.34	0.02	44	1.46

80	Tangipahoa River	42.98	1.28	0.26	0.068162	45.1	
81	Tangipahoa River	31.7	0.76	0.36	0.053227	44	
82	Red River	253.6	0.81	0.48	0.072	45.1	1.20
83	Red River	161.5	0.4	0.34	0.02	44	1.44
84	Red River	152.4	1.62	0.61	0.032	143.8	1.44
85	Red River	155.1	3.96	0.29	0.06	130.5	1.24
86	Red River	248.11	4.82	0.31	0.065235	143.8	
87	Sabine River, LA	116.4	3.66	0.45	0.057	227.6	1.17
88	Sabine River, LA	160.3	1.74	0.47	0.036	177.7	1.17
89	Sabine River, TX	14.2	1.65	0.58	0.054	131.3	2.53
90	Sabine River, TX	12.2	2.32	1.06	0.054	308.9	2.05
91	Sabine River, TX	21.3	0.5	0.13	0.037	12.8	1.47
92	Sabine River, TX	21.64	0.61	0.08	0.04237	12.8	
93	Sabine River, TX	17.37	1.23	0.04	0.050338	14.7	
94	Sabine River, TX	31.39	1.43	0.13	0.041029	24.2	
95	Wind/Bighom rivers	44.2	1.4	0.99	0.14	184.6	1.56
96	Wind/Bighom rivers	85.3	2.4	1.73	0.15	464.6	1.56
97	Copper Creek	16.7	0.5	0.2	0.08	16.8	2.54
98	Clinch River	48.5	1.2	0.21	0.07	14.8	1.25
99	Copper Creek	18.3	0.4	0.15	0.12	20.7	2.54
100	Powell River	36.8	0.9	0.13	0.05	15.5	
101	Clinch River	28.7	0.6	0.35	0.07	10.7	2.20
102	Copper Creek	19.6	0.8	0.49	0.1	20.8	1.14
103	Clinch River	57.9	2.5	0.75	0.1	40.5	
104	Conchelaa Canal	24.7	1.6	0.66	0.04	5.9	1.14
105	Clinch river	33.53	0.78	0.19	0.049483	10.7	
106	Clinch river	55.78	2.26	0.69	0.098768	36.93	
107	Clinch river	53.2	2.4	0.66	0.11	36.9	
108	Coachell canal, CA	23.77	1.6	0.67	0.04	5.96	1.14
109	Coachell canal, CA	24.99	1.54	0.66	0.037	5.92	
110	Copper Creek	16.8	0.5	0.24	0.08	24.6	
111	Missoury river	180.6	3.3	1.62	0.08	1486.5	
112	Bayou Anacoco	25.9	0.9	0.34	0.07	32.5	
113	Bayou Anacoco	36.6	0.9	0.4	0.07	39.5	
114	Nooksack river	64	0.8	0.67	0.27	34.8	1.30
115	Wind/Bighom rivers	59.4	1.1	0.88	0.12	41.8	1.18
116	Wind/Bighom rivers	68.6	2.2	1.55	0.17	162.6	1.18
117	John day river	34.1	2.5	0.82	0.18	65	1.89
118	Yadkin River	70.1	2.4	0.43	0.1	111.5	2.17
119	Yadkin River	71.6	3.8	0.76	0.13	260.1	2.17
120	Colorado River	106.1	6.1	0.79	0.088201	181	

121	Colorado River	71.6	8.2	1.2	0.336784	243	
122	Albert	100	4.4	0.029	0.0016	0.2	
123	Dessel-Herentals	35	2.5	0.037	0.0022	0.2	
124	Yuma Mesa A	7.6	3.45	0.68	0.047	0.5	
125	Bocholt-Dessel	35	2.5	0.107	0.0063	1.4	
126	Villemsvaart	34	2.5	0.13	0.0079	1.7	
127	Chicago Ship Canal	49	8.07	0.27	0.019	3	
128	Irrigation	1.4	0.19	0.38	0.11	9.6	
129	Irrigation	1.5	0.14	0.33	0.1	1.9	
130	Puneha	5	0.28	0.26	0.21	7.2	
131	Kapuni	9	0.3	0.37	0.15	8.4	
132	Kapuni	10	0.35	0.53	0.17	12.4	
133	Manganui	20	0.4	0.19	0.18	6.5	
134	Waiongana	13	0.6	0.48	0.24	6.8	
135	Stony	10	0.63	0.55	0.3	13.5	
136	Waiotapu	11.4	0.75	0.41	0.061	8	
137	Manawatu	59	0.72	0.37	0.07	32	
138	Manawatu	63	1	0.32	0.094	22	
139	Manawatu	60	0.95	0.46	0.092	47	
140	Tarawera	25	1.21	0.73	0.084	27	
141	Tarawera	20	1.92	0.62	0.123	11.5	
142	Tarawera	25	1.38	0.77	0.091	20.5	
143	Tarawera	25	1.4	0.78	0.091	15.5	
144	Tarawera	25	1.57	0.83	0.096	18	
145	Tarawera	85	2.6	0.69	0.06	52	
146	Waikato	120	2	0.64	0.05	67	
147	Miljacka	11	0.29	0.35	0.058	2.7	
148	Upper Tame	9.9	0.83	0.46	0.09	5.5	
149	Upper Tame	9.9	0.92	0.52	0.1	5.1	