A comparative evaluation of shear stress modeling based on machine learning methods in small streams

Onur Genç, Bilal Gonen and Mehmet Ardıçlıoğlu

ABSTRACT

Predicting shear stress distribution has proved to be a critical problem to solve. Hence, the basic objective of this paper is to develop a prediction of shear stress distribution by machine learning algorithms including artificial neural networks, classification and regression tree, generalized linear models. The data set, which is large and feature-rich, is utilized to improve machine learning-based predictive models and extract the most important predictive factors. The 10-fold cross-validation approach was used to determine the performances of prediction methods. The predictive performances of the proposed models were found to be very close to each other. However, the results indicated that the artificial neural network, which has the R value of 0.92 ± 0.03, achieved the best classification performance overall accuracy on the 10-fold holdout sample. The predictions of all machine learning models were well correlated with measurement data.

Key words | ANN, C&R-T, GENLIN, machine learning, shear stress distribution, small streams

Onur Genç (corresponding author)
Department of Civil Engineering,
Melikşah University,
Kayseri,
Turkey
E-mail: *ogenc@meliksah.edu.tr*

Bilal Gonen
Department of Computer Science,
University of West Florida,
Pensacola, FL,
USA

Mehmet Ardıçlıoğlu
Department of Civil Engineering,
Erciyes University,
Kayseri,
Turkey

INTRODUCTION

River flows are affected by turbulent processes and also have complex and three-dimensional structures. Owing to this, to explain the behaviors of flow properties such as discharge, velocity, and shear stress distributions is difficult. The distribution of shear stress in an open channel is influenced by many factors, such as roughness, the structure of the secondary current, existence of free water surface, and the geomorphology of the cross section (Ghosh & Roy 1970; Knight & Patel 1985; Yang & Lim 2005). The shear stress and its distribution have great importance for the estimation of sediment and pollutant transport, river resistance, bank protection, and river management in hydraulic engineering.

The pioneer investigations on shear stress distribution were carried out by Leighly (1932). Recently, much research has been performed in order to determine shear stress distribution using direct and indirect methods. Knight (1981) proposed an empirical equation for the distribution of shear stress along the channel wetted perimeter.

Jin *et al.* (2004) developed a semi-analytical model to predict the boundary shear stress distribution in straight, non-circular ducts and open channels. They reached a satisfactory result to predict the boundary shear along the whole length of the side walls of trapezoidal and rectangular channels. Yang (2010) improved a method for determining shear stress distribution in steady, uniform, and fully developed turbulent flows by applying an order-of-magnitude analysis to the Reynolds equations. He accurately predicted mean shear stress in trapezoidal channels without empirical coefficients involved. Javid & Mohammadi (2012) calculated the average bed and wall shear stress in a straight prismatic trapezoidal channel using Guo and Julien's method. They found $R^2 = 0.99$, and average relative error less than 5.35% in this method.

Bonakdari & Levacher (2010) studied boundary shear stresses for rectangular open channels of different roughness, using computational fluid dynamics (CFD). Yang (2010) established a theoretical relationship between the boundary shear stress and depth-averaged apparent shear stress in open channels.

Machine learning algorithms (MLAs) are an alternative approach to predicting the flow properties in hydraulics

doi: 10.2166/hydro.2015.142

engineering. The input and output variables are selected for the MLA and can be used with modern regression techniques to fit the measured data. MLAs include the training and testing parts to develop models that can be learnt from experience and data (Mitchell 1997).

Bhattacharya *et al.* (2007) modeled the sediment transport using two machine learning approaches (artificial neural networks (ANNs) and model trees). For field measurements, the MLA models outperform the existing models. In addition, the MLA model gives the least errors. They stressed that the utilization of MLA in sediment transport modeling can be proposed, and further research in this area is strongly recommended.

Samandar (2011) investigated the friction coefficient in open channel flow using an adaptive neural-based fuzzy inference system (ANFIS). He showed that there is a good correlation between the experimental data and predicted results. Genc *et al.* (2014) studied the mean velocity and discharge for small streams using ANNs and ANFIS. They compared the accuracy of these models using multiple-linear regression models and found that the ANFIS model performed better than the ANN.

MLA has been applied to flood forecasting by Han *et al.* (2007). They investigated an optimum selection among a large number of various input combinations and parameters for any modelers in using support vector machines (SVM). They made a comparison with some benchmarking models such as transfer function, trend and naive models. They reported that SVM is able to surpass all of these models in the test data, at the expense of a huge amount of time and effort. They also revealed that linear and non-linear kernel functions can yield superior performances against each other under different circumstances in the same catchment (Shrestha *et al.* 2014).

Azamathulla & Jarrett (2013) investigated the utilization of gene-expression programing to estimate the Manning's roughness coefficient for high-gradient streams. The determination of Manning's n values has much importance for researchers and field engineers. They have reached substantially more effective results than the classical methods.

The basic aim of this paper is to investigate the applicability of the MLA approach as a reliable and efficient method to determine the shear stress in small streams. The velocity measurements, which were carried out by the first and third author in central Turkey, were utilized to model the shear stress.

A BRIEF REVIEW OF THE SHEAR STRESS DISTRIBUTION FOR OPEN CHANNELS

Shear stress distribution is related to the shape of the cross section, flow resistance, sediment transport rate, side wall correction, and channel erosion, etc. Shear stress is not always uniformly distributed over the perimeter of the cross section. A simplifying cross section averaging one-dimensional hydraulic equations was preferred to determine the flow properties in small streams. Conventional methods include velocity samples and empirical formulas. Some characteristics, such as the energy line slope and the roughness, tend to vary with time and water depth section by section through the flow direction. Therefore, application of traditional methods is difficult particularly in an unsteady non-uniform flow (Ardiclioglu *et al.* 2012). For uniform flows, average shear stress at a cross section can be given as the following equation:

$$\tau_0 = \gamma RS \tag{1}$$

where τ_0 expresses the shear stress. γ shows the specific gravity of water. R is the hydraulic radius ($=A/P$ in which A is the wetted area and P is wetted perimeter). S is the energy line slope.

Schlicting (1987) proposed another approach that based logarithmic relation between the shear velocity and the variation of velocity with height for local bed shear stress

$$\frac{u}{u*} = \frac{1}{\chi}\ln\left(\frac{z}{k_s/30}\right) \tag{2}$$

where u signifies the velocity at z. z represents the distance from the bottom of the roughness elements. $u*$ indicates the shear velocity, ($=(\tau_0/\rho)^{1/2}$, in which ρ is the water density). χ shows the von Karman constant. k_s is the Nikuradse's original uniform sand grain roughness. When shear velocities, $u*$, are known, mean shear stresses can be calculated for any vertical.

DATA SOURCES AND FIELD MEASUREMENTS

In this paper, the data source that was used to validate the methodology was collected through field measurements in central Turkey. Turkey has a semi-arid climate with some extremities in temperature. Winters are long and cold in Central and Eastern Anatolia, but mild and short in coastal regions. Flow measurements were carried out by a team that consisted of the first and third authors.

The data set was obtained using the acoustic Doppler velocimeter (ADV) in four different small streams. At the stations shown in Figure 1, 22 field measurements were performed to model the shear stress distribution on the Zamantı River, located in the Seyhan basin, and the Kızılırmak River, located in the Kızılırmak basin. Bunyan, Barsama, and Şahsenem stations are on Sarımsaklı Stream, a tributary of the Kızılırmak River which is the longest river in Turkey. Sosun station is on the Sosun Stream, which is a branch of the Zamantı River.

Each data file consists of information about flow properties that were obtained in these stations between the dates 2005 and 2010. It includes measured flow characteristics, which are given in Table 1.

In Table 1, columns 1 and 2, the stations' names, number of measurements, and dates are presented. In column 3, Q denotes the discharges that were determined using the velocity–area method. In columns 4 and 5, the mean velocity, U_m, and the free water surface velocity, u_{ws}, may be seen respectively. The cheapest and easiest way to determine water surface velocity is to simply float something down the stream and see how fast it goes. In field measurements, water surface velocities, u_{ws}, were readily obtained with an object that is movable on the water surface. In this study, a chronometer was used to measure how many seconds it took for a tree branch to pass a distance of 10 meters.

In Table 1, columns 6–9, variables pertaining to the shape of the measured cross section are presented. A is the cross section area. $H_{\max}$ is the maximum water depth. T is the water surface width. T/R is the aspect ratio, with R ($=A/P$) being the hydraulic radius where P is the wetted perimeter. As seen in column 10, S_{ws} is the water surface slope. In column 11, Re ($=4U_mR/\upsilon$) is the Reynolds number, with υ being the kinematic viscosity. In column 12, Fr ($=U_m/(gH_{max})^{1/2}$) is the Froude number where g is the gravitational acceleration. According to the water surface width, T cross sections were divided by number of slices, n, for each flow condition.

According to the measurement data, these stations are relatively small with shallow streams, where the maximum water depths range between 0.26 and 0.86 m in the measured cross sections. The water surface width varies between values of 2.3 and 9.0 m. Re numbers vary between the values 0.32×10^6 and 1.47×10^6 and Froude number are between the values 0.084 and 0.578. When considering the Re and Fr numbers, these stations have turbulent flow conditions.

Recent studies showed that the ADV device is well-suited to measure turbulent velocities in small streams; therefore, it is assumed that the velocity signal outputs are

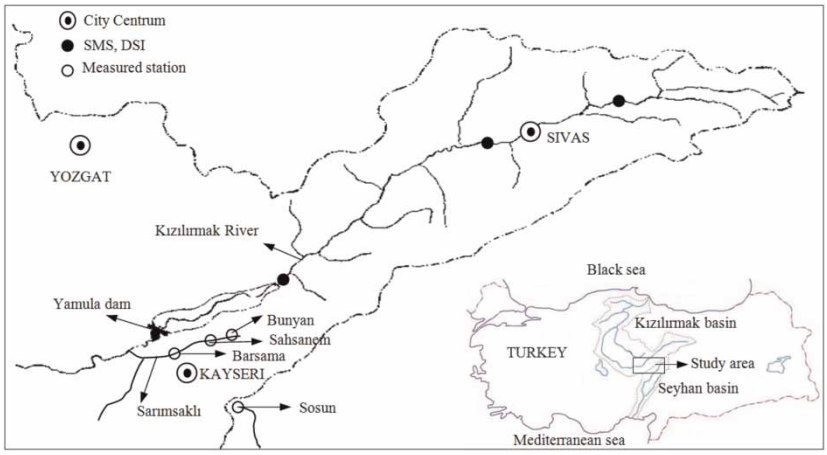

Figure 1 | Location of the study area and measurement stations at Bünyan, Barsama, Şahsenem, and Sosun (Ardiclioglu *et al.* 2012).

Table 1 | Main flow characteristics for measurements

Stations	Dates d/m/y	Q m^3/s	U_m m/s	u_{ws} m/s	A m^2	H_{max} m	T m	T/R	S_{ws}	Re ($\times 10^6$)	Fr	n
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Barsama_1	28/05/2005	1.810	0.890	1.600	2.23	0.39	8.3	34.00	0.0091	0.76	0.481	7
Barsama_2	19/05/2006	2.440	1.051	1.850	2.03	0.40	9.0	35.20	0.0036	0.94	0.531	7
Barsama_3	19/05/2009	3.930	1.214	2.080	2.11	0.45	9.0	29.70	0.0094	1.47	0.578	9
Barsama_4	31/05/2009	0.970	0.590	1.140	2.67	0.26	8.4	45.40	0.0092	0.40	0.333	8
Barsama_5	24/03/2010	1.510	0.806	1.550	2.79	0.38	8.6	34.40	0.0097	0.61	0.417	4
Barsama_6	18/04/2010	2.150	0.865	1.630	2.48	0.38	8.8	22.10	0.0120	0.85	0.421	5
Şahsenem_1	29/03/2006	0.816	0.354	1.040	2.04	0.72	6.0	26.80	0.0059	0.47	0.350	5
Şahsenem_2	20/10/2007	0.718	0.214	0.930	2.32	0.66	5.4	21.90	0.0061	0.46	0.298	9
Şahsenem_3	22/03/2008	0.792	0.301	0.800	3.24	0.72	6.0	22.10	0.0037	0.49	0.314	9
Şahsenem_4	03/05/2008	0.613	0.405	1.000	1.64	0.85	5.4	25.10	0.0045	0.39	0.307	9
Şahsenem_5	11/10/2008	0.667	0.426	1.010	1.87	0.86	5.5	22.00	0.0046	0.44	0.303	9
Şahsenem_6	08/11/2008	0.732	0.286	1.000	2.48	0.79	5.6	19.60	0.0064	0.51	0.282	10
Bünyan_1	24/06/2009	0.788	0.600	0.650	1.40	0.28	4.0	7.00	0.0020	0.71	0.133	7
Bünyan_2	08/02/2010	0.434	0.529	0.400	1.36	0.32	4.0	7.50	0.0030	0.40	0.084	7
Bünyan_3	27/09/2009	0.636	0.565	0.540	1.40	0.33	3.9	8.20	0.0022	0.50	0.113	6
Bünyan_4	04/04/2010	1.082	0.518	0.740	1.18	0.32	4.0	7.30	0.0018	0.78	0.140	4
Bünyan_5	16/05/2010	1.188	0.536	0.540	1.24	0.32	4.0	7.00	0.0024	0.85	0.147	4
Bünyan_6	20/06/2010	0.708	0.516	0.530	1.40	0.34	3.9	7.30	0.0010	0.53	0.103	4
Sosun_1	19/05/2009	0.886	0.561	0.960	1.58	0.62	3.2	7.49	0.0032	0.84	0.227	6
Sosun_2	31/05/2009	0.294	0.285	0.630	1.03	0.43	3.0	9.49	0.0016	0.32	0.144	5
Sosun_3	24/03/2010	0.338	0.327	0.630	1.03	0.45	2.9	8.85	0.0026	0.37	0.156	5
Sosun_4	18/04/2010	0.529	0.541	0.930	0.98	0.54	2.3	6.53	0.0034	0.67	0.235	5

'true' turbulent velocity data (Chanson *et al.* 2008). For our measurements, SonTek Flow Tracker was preferred. Some technical characteristics of ADV are given as follows by SonTek handheld: velocity range, ±0.001–4.5 m/s; velocity resolution, 0.0001 m/s; velocity accuracy, ±1% of measured velocity; operating temperature, −20–50 °C (SonTek 2002). ADV is designed to record instantaneous three-dimensional (u, v, w) velocity components in a sampling volume using Doppler shift effect (Zedel *et al.* 1996; Nikora & Goring 1998). ADV can measure and record velocity samples by sending out short acoustic waves from the transmitter probe. Point velocities were measured in the vertical direction starting from a point that is 4 cm above the streambed for each vertical slice. The same procedure was repeated every 2 cm from this point to the water surface for each vertical slice. Meanwhile, the ADV remained in a fixed position in the stream. The probe position was adjusted manually vertical by vertical. An illustrative photo of site surveying and flow measurement at the Şahsenem station is shown in Figure 2.

Figure 2 | Velocity measurements at Şahsenem Station (Ardiclioglu *et al.* 2012).

In these field measurements, each vertical velocity distribution along the cross section at four stations was determined. As mentioned above, according to the water surface width, the number of verticals at each station measured cross section was decided. The number of verticals varies from 4 up to 10. Afterwards, shear stress distribution in the measured vertical was calculated using these vertical velocity distributions. The shear velocity, u_*, and roughness parameter, k_s, can be determined by Equation (2) using the von Karman constant, $1/\chi = 25$, proposed by Sümer (2004) for given measured velocity profile $u(z)$. The point velocities, u, against z are plotted in semi-log graphs. The $0.1H \leq z \leq (0.2\text{–}0.3)H$ interval shows where the logarithmic layer is supposed to lie, as given in Figure 3. The z axis shows the water depth, H, at a measured vertical. Extending the straight line portion of the velocity distribution finds its z-intercept, and this is equal to $k_s/30$. Using $k_s/30$ values and shear velocities, u_*, having the best fit with measured data can be determined by Equation (2).

A sample vertical velocity distribution can be seen in Figure 3 for Şahsenem_2 at $y = 170$ cm. First, the measured vertical velocities, u, were plotted against H in a semi-log graph, and the logarithmic layer was determined for measured vertical slices, as seen in Figure 3(b). Using this straight line, $k_s/30$ value was obtained as 1.85 (in bold) in Table 2. Then, shear velocities u_* having the best fit with measured data were determined using Equation (2) by trial and error method.

Figure 3(a) shows that the best fit velocity distributions were obtained for measured data by estimated shear velocities as 0.12 m/s (in bold). For each measurement, shear velocities, u_*, were calculated using vertical measured velocities for the cross section. Shear velocities have been calculated for all measurements. When shear velocities u_* were known, shear stress distribution can be calculated by $\tau_{0=}(u_*)^2\rho$ in measured vertical slices. The specific weight of water was used as $\rho = 1000$ kg/m^3.

Similar studies were done for all flow conditions at the stations. In Table 2, the obtained roughness parameter, $k_s/30$, is given in line 1, obtained shear velocities, u_*, in line 2, and lateral measured point distance from channel wall through measurement direction, y, in line 3. Calculated shear stress in the measured vertical is presented in line 4. Also, calculated shear stress distribution is demonstrated in Figure 4. It is seen that shear stress distribution in the middle of the cross section increases and close to the sidewall decreases. Similar studies were done for all flow conditions at stations.

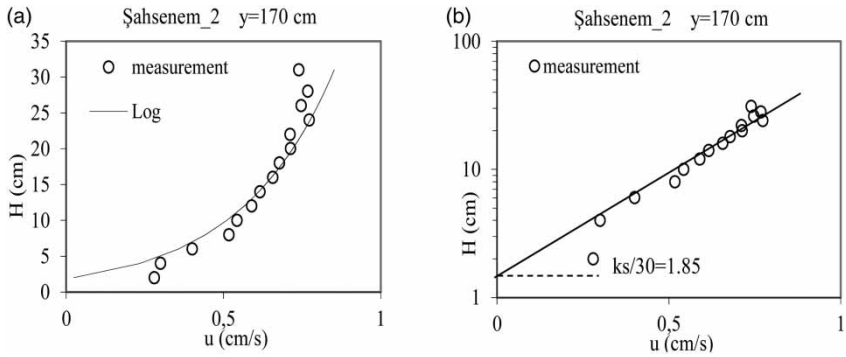

Figure 3 | Velocity distribution for Şahsenem_2, $y = 170$ cm.

Table 2 | Parameters to calculate the shear stress distribution at Şahsenem_2 station

y (cm)	70	120	170	220	270	320	370	420	470
$k_s/30$	1.95	1.20	**1.85**	1.60	1.60	1.40	1.90	1.90	2.00
u_* (m/s)	0.09	0.10	**0.12**	0.11	0.13	0.10	0.12	0.10	0.09
τ_0 (N/m^2)	8.46	10.40	13.46	12.54	15.88	10.00	14.40	9.12	7.23

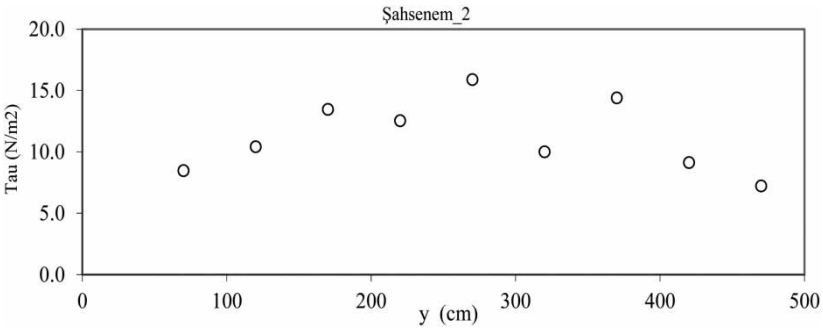

Figure 4 | Calculated shear stress distribution for Şahsenem_2 station.

PREDICTION MODELS

Predictive modeling is the process by which a model is created or chosen to try to best predict the probability of an outcome (Geisser 1993). In many cases, the model is chosen to guess the probability of an outcome given a set amount of input data, not unlike an email determining how likely it is that it is spam. Predictive analytics is a broad term describing a variety of statistical and analytical techniques used to develop models that predict future events or behaviors. The form of these predictive models varies, depending on the behavior or event that they are predicting. Most predictive models generate a score (a credit score, for example), with a higher score indicating a higher likelihood of the given behavior or event occurring (Nyce 2007).

ANNs

ANNs are one of the most popular machine learning methods. ANN has the capability of learning the mathematical correlation between input and output of nonlinear systems. The concept of ANN comes from the operation of neurons in the human brain. An artificial neuron is an engineering approach of a biological neuron. It has multiple inputs and one output. ANN consists of a large number of simple processing elements that are interconnected with each other and layered (Li 1994; Christos & Siganos 2010). ANN has artificial neurons and these neurons receive inputs from other elements. These inputs are weighted and summed. After this operation, the result is transformed by a transfer function into the output.

Neurons are connected in an organized way and form a neural network. In the feed-forward neural networks, also known as multilayer perceptrons, the neurons are in layers (Figure 5). Usually, there is one layer as an input layer, one layer as an output layer, and between the input and output layers are the hidden layers. Hidden layers may be one or more layers. Each layer is fully connected with the layers before and after them. There are weights associated with the connections that go from one neuron to another neuron. These weights represent the strength of influence between the neurons. To make the predictions, information goes from an input layer, passes through the hidden layers, and finally reaches the output layer (IBM SPSS Modeler 2012).

Classification and regression tree

The classification and regression tree (C&R-T) model is a tree-based characterization and prediction method. This method recursively divides the training set into similar parts. The C&R model examines the input fields to determine the best split scenario. At the end of each split, the training records are split into two subgroups. This is a recursive process, and this recursion continues until a stopping criterion is met.

The purpose of creating the C&R-T is to have subgroups with similar output values. This similarity is measured by some type of node impurity measure. The split is made only if the split for a branch reduces the impurity by less than a predetermined value. The least-square deviation criterion is used to calculate the level of impurity for regression-type problems. There are two fields in the data set: frequency field and case weight field. These two fields are used to reduce the size of the data set. The frequency field means the number of observations that each record represents. It is useful to reduce the size of the data set, because instead of having one record for every individual, one record

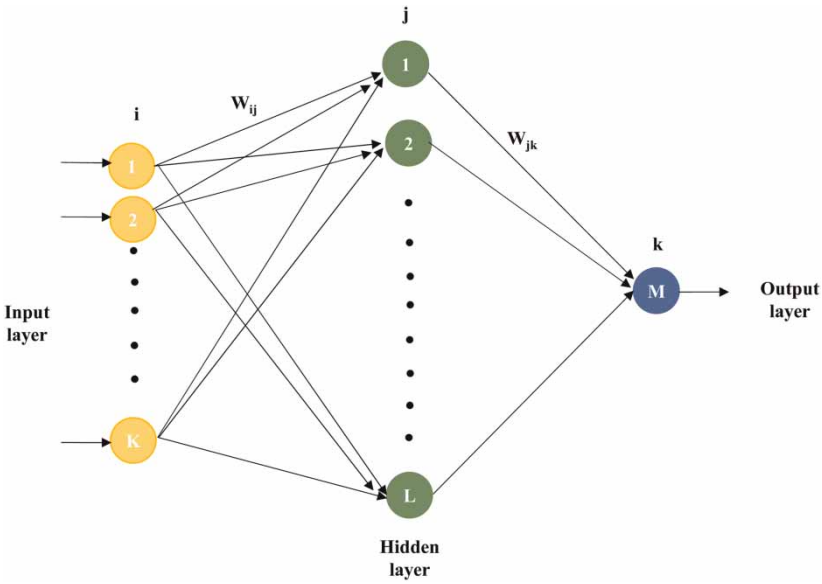

Figure 5 | A three-layer ANN architecture.

can represent multiple individuals. The total number of observations in the sample is the same as the sum of the values for the frequency field.

The case weight field is used when the records in the data set are to be treated unequally. This unequal treatment helps to reduce the size of the data set. Consider a survey made in a school, where 100 students respond and 10,000 students do not respond. If you define a case weight equal to 1 for responders and 100 for non-responders, then you can include all of the responders but just 1% of the non-responders. This way, the size of the data file can be considerably reduced (IBM SPSS Modeler 2012).

Generalized linear models

Generalized linear models (GENLINs) are used in many areas of prediction such as in regression and classification as well. It makes it possible to look for linear and non-linear relationships between a continuous, or binomial, multinomial categorical dependent variable and categorical or continuous predictor variables. This approach is used when the normality and constant variance assumptions are not satisfied. A number of widely used types of analysis can be considered as special applications of GENLIN, such as binomial and multinomial logit and prohibit regression models. A GENLIN usually makes the distribution assumptions that the response variable is independent and can have any distribution from an exponential density family.

Many widely used statistical models belong to GENLIN. For example, classical linear models with normal errors, logistic and prohibit models for binary data, log-linear models for multinomial data, poisson, binomial, gamma and normal distribution, etc. These can be formulated as a GENLIN by selecting an appropriate link function and a response probability distribution. If the identity function is chosen as the link along with the normal distribution, then ordinary linear models are recovered as a special case (Belgin 2010). GENLINs are an extension of linear regression models and consist of several components:

1. A dependent variable z whose distribution is parameter Q.
2. A set of independent variables $x_1, \ldots, x_m$ and predicted $Y = \sum_{i=1}^{m} \beta_i x_i$.
3. A linking function $Q = f(Y)$ connecting the parameter Q of the distribution of z with the Y of the linear model.

When z is normally distributed with mean Q and variance and when $Q = Y$, we have ordinary linear models with normal errors (McCullagh *et al.* 1989).

K-fold cross-validation

The problem of selecting the best algorithm arises in several cases (Rice 1975). Cross-validation is an accuracy estimation method. Estimating the accuracy of a classifier induced by supervised learning algorithms is important, not only to predict its future prediction accuracy, but also for choosing a classifier from a given set (model selection), or combining classifiers (Kohavi & John 1997). Cross-validation is a popular strategy for algorithm selection. The main idea behind cross-validation is to split data, once or several times, for estimating the risk of each algorithm. Part of the data (the training sample) is used for training each algorithm and the remaining part (the validation sample) is used for estimating the risk of the algorithm. Then, cross-validation selects the algorithm with the smallest estimated risk (Arlot & Celisse 2010).

Performance criteria

Root mean square errors (RMSE), mean absolute errors (MAE), residual mean error (RME), and correlation coefficient, R statistics were used as performance criteria to compare our prediction models. These performance criteria are presented, respectively, as the following forms:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_{i\text{-}observed}} - \mathrm{y_{i\text{-}estimated}})^2} \quad (3)$$

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}|\mathrm{y_{i\text{-}observed}} - \mathrm{y_{i\text{-}estimated}}| \quad (4)$$

$$\mathrm{RME} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_{i\text{-}observed}} - \mathrm{y_{i\text{-}estimated}}) \quad (5)$$

In Equations (3)–(5), N is the number of data sets. $\mathrm{y_{i\text{-}observed}}$ denotes the target variable, and $\mathrm{y_{i\text{-}estimated}}$ denotes the predicted value by the model. Correlation coefficient, R, can be determined as in Equation (6):

$$R = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \quad (6)$$

In Equation (6), R signifies the sample correlation coefficient. n is the sample size. x is the value of the independent variable. y is the value of the dependent variable. The higher the R values indicate a better performance for compared models (Everitt & Skrondal 2002).

Sensitivity analysis for models

Sensitivity analysis is a technique which is used to determine how an independent variable impacts a certain dependent variable under some assumptions. Sensitivity analysis is a way to predict the outcome of a decision if a situation turns out to be different compared to the original prediction. Sensitivity analysis can be used to test the robustness of the results of a model or system in the presence of uncertainty. It can be used to increase the understanding of the relationships between input and output variables in a system or model. After redundant parts of the model structures are identified, they can be removed to simplify the model.

CASE STUDY AND DISCUSSION

To demonstrate the suggested methodology in the section 'Prediction models', a most popular data mining toolkit was utilized, namely IBM SPSS Modeler 16. The proposed methodology could be applied for determination of shear stress distribution in small streams. The preliminary analysis showed that ANN, C&R-T, and GENLIN are the most satisfactory models (in terms of the presented performance measures) in predicting the target value, shear stress distribution. The target variable shear stress distribution in measured vertical, τ_0, was predicted through eight observational variables and calculated non-dimensional parameters. These parameters which symbolized y/T, z/h, T/H, z/T, z/y, T/R, S_{ws}, and u_{ws}, were defined in the previous sections. The number of 145 shear stresses, τ_0, that were calculated using Equation (2) for each vertical in the measured cross sections were modeled in this paper using the proposed methodology. The results which were obtained by also employing 10-fold cross-validation for each method are tabulated in terms of discussed criteria metrics (linear correlation, R, and mean squared error (MSE)) in Table 3. ANN has the R value of 0.92 ± 0.03 with the average MSE value of 4.89 ± 1.79. It is

Table 3 | Ten-fold cross-validation results for the machine learning prediction models

	ANNs		CR-T		GENLIN	
Fold	MAE (N/m^2)	*R*	MAE (N/m^2)	*R*	MAE (N/m^2)	*R*
1-fold	7.10	0.92	7.17	0.85	6.00	0.94
2-fold	6.33	0.94	7.38	0.93	5.46	0.93
3-fold	7.15	0.88	7.91	0.93	7.37	0.89
4-fold	3.92	0.95	4.32	0.88	3.85	0.94
5-fold	3.66	0.88	3.15	0.91	**2.87**	**0.95**
6-fold	**2.09**	**0.96**	3.04	0.80	4.95	0.85
7-fold	4.63	0.92	4.90	0.93	5.55	0.89
8-fold	2.66	0.86	**2.78**	0.86	3.85	0.84
9-fold	5.88	0.94	5.34	0.94	6.23	0.88
10-fold	5.48	0.93	3.46	**0.97**	5.13	0.94
Mean	4.89	0.92	4.95	0.90	5.13	0.90
SD	1.79	0.03	1.94	0.05	1.32	0.04

commonly accepted that if *R* is higher than 0.8, the predictive model has performed fairly well (Hair *et al.* 1998). All our models have passed this threshold. The obtained results reveal that ANNs outperformed the two other powerful machine learning algorithms (e.g., C&R-T and GENLIN). The lowest values of MAE, and the highest values of *R* for each model and fold are shown in bold in Table 3.

Application of ANN model

In our artificial neural network model, we employed the multilayer perceptron (MLP) type of network algorithms with one hidden layer that has between 5 and 10 neurons to predict the target variable, since this specific combination has provided higher results in our preliminary analysis. First of all, the network architecture and ANN model were generated using all the eight input variables which are explained in the previous section. Then, as seen in Table 4, four performance criteria explained in the previous section were performed to evaluate our ANN model. The average values of RMSE, MAE, RME, and *R* were calculated as 7.53 N/m^2, 4.89 N/m^2, 0.27 N/m^2 mm, and 0.92 N/m^2, respectively, for the ANN model. Our ANN model has showed slightly better performance than the other models.

Determination of the contribution of each predictor in predicting the shear stress the sensitivity analysis procedure defined previously was accepted. The ranking of the

Table 4 | The inputs and performance indices, RMSE, MAE, RME, and *R* statistics for each model

Models	Inputs	RMSE (N/m^2)	MAE (N/m^2)	RME (N/m^2)	*R*
ANN	y/T, z/h, T/H, z/T, z/y, T/R, S_{ws}, and u_{ws}	7.53	4.89	0.27	0.92
C&R-T	y/T, z/h, z/y, T/R, S_{ws}, and u_{ws}	7.82	4.95	0.12	0.90
GENLIN	y/T, z/h, T/H, z/T, z/y, T/R, S_{ws}, and u_{ws}	8.47	5.13	− 0.31	0.90

normalized predictor variables' importance was calculated during the testing studies and demonstrated in Figure 6. As shown in Figure 6(a), T/H, T/R, and S_{ws} are the most important variables for our ANN model. T/H has a predictor importance of 0.23 and is better than T/R, which has a predictor importance of 0.22, with small differences. S_{ws} has a predictor importance of 0.16.

When all the pairs of predicted and observed shear stress, τ_0, from all measured stations are drawn on the same figure, we obtain the linear relationship between the predicted and observed shear stress, τ_0, in Figure 7. It is clear from the figure that the predictions of the ANN are less scattered and closer to the exact line (45 °) than the C&R-T and GENLIN models with small difference. Particularly when the observed shear stress values are between 0 and 20 N/m^2, the predicted and observed values are quite consistent. As seen in Table 3, for our ANN model, minimum MAE and maximum *R* values are in fold 6. The data set in fold 6 is between 0.630 and 19.60 N/m^2. When the observed shear stress values are bigger than 20 N/m^2, predicted values start to deviate from the exact line (45 °).

Application of C&R-T model

For C&R-T, a single, standard model was generated to determine the relationships between fields using six input parameters (Table 4). The standard models are easier to interpret. The tree depth for the current node is the maximum tree depth. As default, it was selected as 5. There are three different impurity measures (Gini, twoing, and the least-squared deviation) used to find splits for C&RT models. Gini index was utilized to find splits in this study.

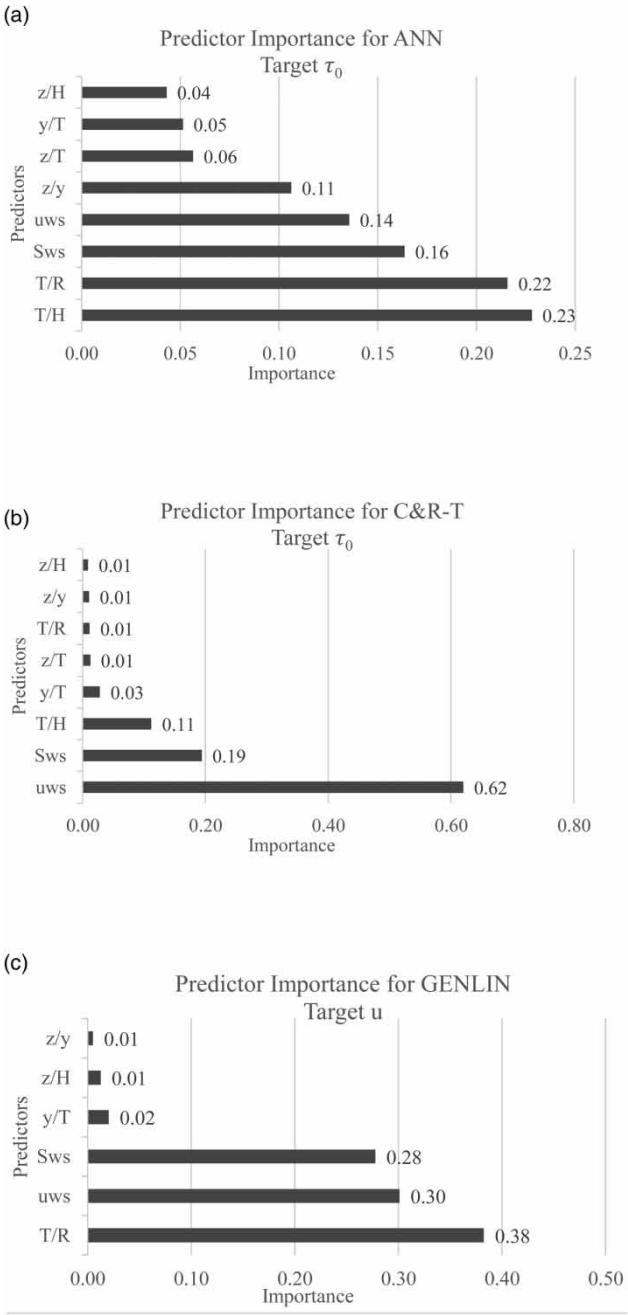

Figure 6 | The predictor importance for machine learning models.

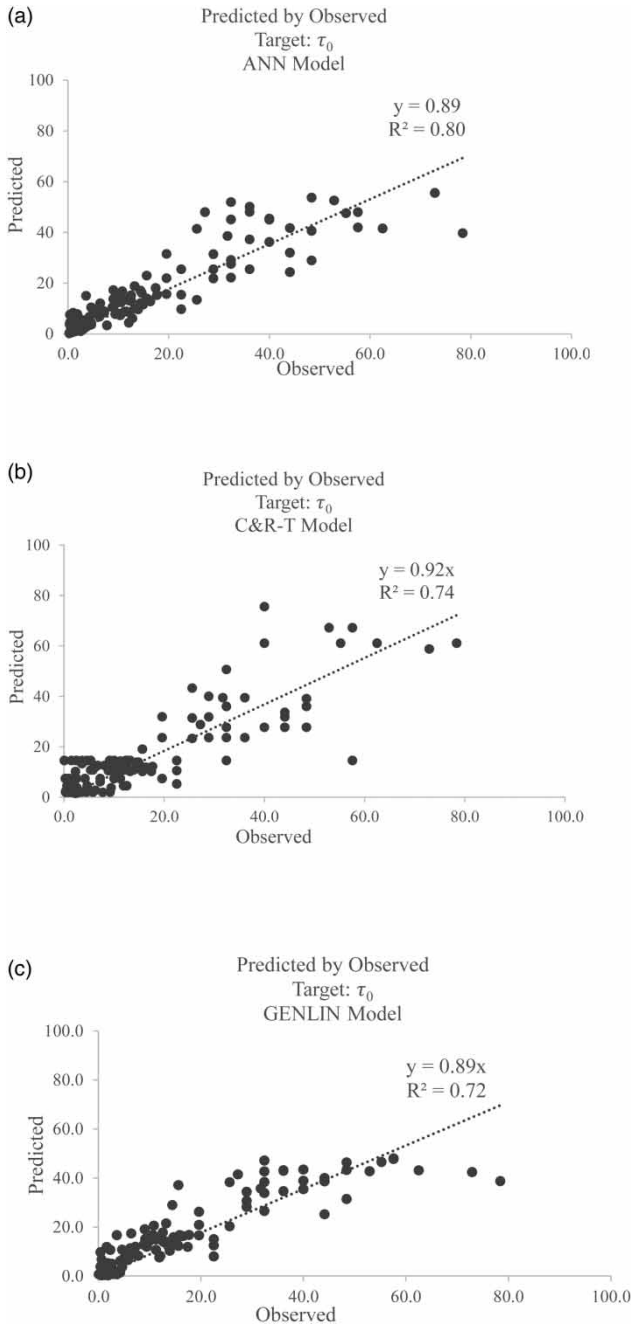

Figure 7 | The scatter plots of the observed and predicted values by ANN, C&R-T, and GENLIN models.

As given in Table 3, our C&RT model has the mean R value of 0.90 ± 0.05 with the mean MSE value of 4.95 ± 1.32 for each fold. For our C&RT model, minimum MAE is 2.78 in fold 8 and maximum R value is 0.97 in fold 10. The data set in fold 10 is between 0.40 and 57.60 N/m^2. C&R-T has the RMSE and RME values of 7.82 and 0.12 N/m^2 as shown in Table 4. As presented in Figure 6(b), u_{ws} is the most important variable that has an importance of 0.62 in our C&R-T model. The variables S_{ws} and T/H have the values of 0.19 and 0.11, respectively. As seen in the graphs in Figure 7(b), it can be understood that the predictions of the C&R-T model are less scattered and close to the exact line (45 °). As in the

ANN model, when the observed shear stress values are bigger than 20.0 N/m^2, they start to deviate from the exact line (45 °). The tree shape and the mean values for fold 5 (predictions) are demonstrated in Table 5. As seen in the table, the first branch of the tree is u_{ws} which has the highest importance at 0.70 for fold 5 and the critical u_{ws} value is 1.087 m/s. The second branch of the tree, T/H, has the second highest importance as 0.13 for fold 5. Its critical value is 13.341.

Application of GENLIN model

All the eight inputs were used to predict the target value, shear stresses, τ_0 in our GENLIN model. There are six different log-likelihood functions such as normal, inverse Gaussian, gamma, negative binomial, poisson, and binomial (m) for probability distribution in the GENLIN model. Normal distribution was selected in this study. Canonical and default link functions for probability distributions were tabulated in the GENLIN model guide. Identity link function as canonical link was determined for normal distribution in the GENLIN model. Several statistics (deviance, Pearson chi-square, maximum likelihood estimate, and fixed value) are calculated to evaluate goodness-of-fit of a presented GENLIN model. The maximum likelihood estimate as scale parameter method was performed in this study.

As shown in Table 3, the GENLIN model has the mean R value of 0.90 ± 0.04 with the mean MAE value of 5.13 ± 1.32 for each fold. In our GENLIN model, minimum MAE is 2.87 and maximum R value is 0.95 in fold 5. The data set in fold 5 is between 0.90 and 32.40 N/m^2. GENLIN has the RMSE and RME values of 8.47 and −0.31 N/m^2, as shown in Table 4.

Table 5 | Tree shape and the means of the predicted values for fold 5 in C&R-T model

$u_{ws} \leq 1.087$ [Ave: 6.235, Effect: −7.749]
$T/H \leq 13.341$ [Ave: 2.186, Effect: −4.049] ≥ 2.186
$T/H > 13.341$ [Ave: 10.165, Effect: 3.93]
$y/T \leq 0.798$ [Ave: 11.933, Effect: 1.768] ≥ 11.933
$y/T > 0.798$ [Ave: 3.344, Effect: −6.821] ≥ 3.344
$u_{ws} > 1.087$ [Ave: 37.583, Effect: 23.599]
$S_{ws} \leq 0.010$ [Ave: 30.989, Effect: −6.594]
$T/H \leq 27.363$ [Ave: 40.091, Effect: 9.102] ≥ 40.091
$T/H > 27.363$ [Ave: 23.708, Effect: −7.281] ≥ 23.708
$S_{ws} > 0.010$ [Ave: 67.256, Effect: 29.673] ≥ 67.256

As presented in Figure 6(c), T/R is the most important variable that has an importance of 0.38 in our GENLIN model. The variables u_{ws} and S_{ws} have the values of 0.30 and 0.28, respectively. As shown in Figure 6(c), it can be seen that the predictions of the GENLIN model are less scattered and close to the exact line (45 °). As in other proposed models, when the observed shear stress values are bigger than 20.0 N/m^2, they start to deviate from the exact line (45 °).

CONCLUSION

Shear stress distribution is a precious parameter for the investigations of turbulence, sediment transport, and river management. Determining shear stress distribution has been considered a serious problem. This study demonstrates that machine learning-based methodology can be performed to predict the shear stress distribution in small streams. ANN, C&R-T, and GENLIN were indicated to be the best by the preliminary studies implemented to obtain which models perform better. In this study, the eight parameters were utilized as inputs to the models for predicting the shear stress distribution in measured verticals. Importance of predictor variables for these models were revealed. The performances of prediction methods were calculated using the 10-fold cross-validation approach. Our ANN model, which has the R value of 0.92 ± 0.03 with the RMSE value of 7.53 N/m^2, performed better than the other models in predicting the shear stress. In fact, the results that C&R-T and GENLIN are worse have not been presented. It should be particularly expressed that in cases where the shear stress is between values of 0–20 N/m^2, the estimated and observed values are quite consistent for all proposed machine learning models. As mentioned in previous sections, there are small differences between the models. Consequently, it is expressed that all these methods may improve a better understanding of shear stress and its distribution in small streams.

ACKNOWLEDGEMENT

The corresponding author would like to thank The Scientific and Technological Research Council of Turkey (TUBITAK),

since this paper was written during the period of time in which he was supported by this council to pursue a 1-year visiting scholarship program at Auburn University, Auburn, Alabama, USA.

REFERENCES

Ardiclioglu, M., Genc, O., Kalin, L. & Agiralioglu, N. 2012 Investigation of flow properties in natural streams using the entropy concept. *Water Environ. J.* **26**, 147–154.

Arlot, S. & Celisse, A. 2010 A survey of cross-validation procedures for model selection. *Stat. Surveys* **4**, 40–79.

Azamathulla, H. Md. & Jarrett, R. D. 2013 Use of gene-expression programming to estimate Manning's roughness coefficient for high gradient streams. *Water Resour. Manage.* **27** (3), 715–729.

Belgin, K. 2010 *Parameter Estimation in Generalized Partial Linear Models with Tikhonov Regularization*. Dissertation MSc. Thesis, Institute of Applied Mathematics of METU, Ankara, Turkey.

Bhattacharya, B., Price, R. K. & Solomatine, D. P. 2007 Machine learning approach to modeling sediment transport. *J. Hydrol Eng.* **133** (4), 440–450.

Bonakdari, H. & Levacher, M. D. 2010 Numerical study of boundary shear stress distribution in rectangular open channel flow. *XIèmes Journées Nationales Génie Côtier–Génie* **155**, 22–25.

Chanson, H., Trevethan, M. & Aoki, S. 2008 Acoustic Doppler velocimeter (ADV) in small estuary: field experience and signal post-processing. *Flow Measure. Instrument.* **19** (5), 307–313.

Christos, S. & Siganos, D. 2010 Neural Networks. http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

Everitt, B. S. & Skrondal, A. 2002 *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, UK.

Geisser, S. 1993 *Predictive Inference*. Vol. 55. CRC Press, Boca Raton, FL, USA.

Genc, O., Kisi, O. & Ardiclioglu, M. 2014 Determination of mean velocity and discharge in natural streams using neuro-fuzzy and neural network approaches. *Water Resour. Manage.* **28**, 2387–2400.

Ghosh, S. N. & Roy, N. 1970 Boundary shear distribution in open channel flow. *J. Hydr. Div.* **96** (4), 967–994.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. 1998 *Multivariate Data Analysis*. Prentice Hall, Upper Saddle River, NJ, USA.

Han, D., Chan, L. & Zhu, N. 2007 Flood forecasting using support vector machines. *J. Hydroinformat.* **9**, 267–276.

IBM SPSS Modeler 2012 14.2 User's Guide.

Javid, S. & Mohammadi, M. 2012 Boundary shear stress in a trapezoidal channel. *Int. J. Eng.-Trans. A* **25** (4), 323–332.

Jin, Y. C., Zarrati, A. R. & Zheng, Y. 2004 Boundary shear distribution in straight ducts and open channels. *J. Hydraulic Eng.* **130** (9), 924–928.

Knight, D. W. 1981 Boundary shear in smooth and rough channels. *J. Hydr. Div.* **107** (7), 839–851.

Knight, D. W. & Patel, H. S. 1985 Boundary shear stress distributions in rectangular duct flow. In: *Proc. 2nd Int. Symposium on Refined Flow Modelling and Turbulence Measurements*. Iowa, USA.

Kohavi, R. & John, G. H. 1997 Wrappers for feature subset selection. *Artif. Intell.* **97** (1), 273–324.

Leighly, J. B. 1932 Toward a theory of the morphologic significance of turbulence in the flow of water in streams. *Univ. Calif. Publ. Geogr.* **6** (1), 1–22.

Li, E. Y. 1994 Artificial neural networks and their business applications. *J. Inform. Manage.* **27**, 303–313.

McCullagh, P., Nelder, J. A. & McCullagh, P. 1989 *Generalized Linear Models*. vol. 2, Chapman and Hall, London.

Mitchell, T. M. 1997 *Machine Learning*. McGraw-Hill, New York, USA.

Nikora, V. I. & Goring, D. G. 1998 ADV Measurements of turbulence: can we improve their interpretation. *J. Hydraul. Eng.* **124** (6), 630–634.

Nyce, C. 2007 Predictive Analytics White Paper, American Institute for CPCU. Insurance Institute of America, pp. 9–10.

Rice, J. R. 1975 *The Algorithm Selection Problem*. Computer Science Technical Reports, report no. 75-153, Purdue University.

Samandar, A. 2011 A model of adaptive neural-based fuzzy inference system (ANFIS) for prediction of friction coefficient in open channel flow. *Sci. Res. Essays* **6**, 1020–1027.

Schlicting, H. 1987 *Boundary Layer Theory*, 7th edn. McGraw-Hill, New York, USA.

Shrestha, D. L., Kayastha, N., Solomatine, D. & Price, R. 2014 Encapsulation of parametric uncertainty statistics by various predictive machine learning models: MLUE method. *J. Hydroinform.* **16** (1), 95–113.

SonTek 2002 Flow Tracker Handheld ADV, Technical Document.

Sumer, B. M. 2004 *Lecture Notes on Turbulence*. Technical University of Denmark, 2800 Lyngby, Denmark.

Yang, S. Q. 2010 Depth-averaged shear stress and velocity in open-channel flows. *J. Hydraul. Eng.* **136** (11), 952–958.

Yang, S. Q. & Lim, S. Y. 2005 Boundary shear stress distributions in trapezoidal channels. *J. Hydraulic Res.* **43** (1), 98–102.

Zedel, L., Hay, A. A., Cabrera, R. & Lohrmann, A. 1996 Performance of a single-beam pulse-to-pulse coherent Doppler profiler. *IEEE J. Ocean. Eng.* **21** (3), 290–297.

First received 19 December 2014; accepted in revised form 13 March 2015. Available online 28 April 2015