Water Quality Research Journal

Water Quality Research Journal Vol 56 No 3, 127 doi: 10.2166/wqrj.2021.003

A genetic algorithm-based support vector machine to estimate the transverse mixing coefficient in streams

Hosein Nezaratian[a], Javad Zahiri[b], Mohammad Fatehi Peykani[c], AmirHamzeh Haghiabi[d] and Abbas Parsaie[e,*]
[a] Faculty of Engineering and Applied Science, University of Regina, Regina, Canada
[b] Department of Water Engineering, Agricultural Sciences and Natural Resources University of Khuzestan, Mollasani, Khuzestan, Iran
[c] Department of Industrial Engineering, University of Eyvanekey, Semnan, Iran
[d] Water Engineering Department, Lorestan University, Khorramabad, Iran
[e] Hydro-Structure Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran
*Corresponding author. E-mail: abbas_parsaie@yahoo.com

AP, 0000-0001-7312-0634

ABSTRACT

Transverse mixing coefficient (TMC) is known as one of the most effective parameters in the two-dimensional simulation of water pollution, and increasing the accuracy of estimating this coefficient will improve the modeling process. In the present study, genetic algorithm (GA)-based support vector machine (SVM) was used to estimate TMC in streams. There are three principal parameters in SVM which need to be adjusted during the estimating procedure. GA helps SVM and optimizes these three parameters automatically in the best way. The accuracy of the SVM and GA-SVM algorithms along with previous models were discussed in TMC estimation by using a wide range of hydraulic and geometrical data from field and laboratory experiments. According to statistical analysis, the performance of the mentioned models in both straight and meandering streams was more accurate than the regression-based models. Sensitivity analysis showed that the accuracy of the GA-SVM algorithm in TMC estimation significantly correlated with the number of input parameters. Eliminating the uncorrelated parameters and reducing the number of input parameters will reduce the complexity of the problem and improve the TMC estimation by GA-SVM.

Key words: GA-SVM algorithm, pollution, sensitivity analysis, transverse mixing coefficient

HIGHLIGHTS

- Genetic algorithm (GA)-based support vector machine (SVM) was used to estimate TMC in streams.
- Sensitivity analysis showed that the accuracy of GA-SVM algorithm in TMC estimation significantly correlated with the number of input parameters.

1. INTRODUCTION

Increasing the accuracy of modeling the process of pollution release into streams will increase the ability to control the quality of streams and thereby reduce environmental damage. Therefore, the capability to estimate the transport of pollutants in streams and waterways has always been a considerable issue in many industrial and environmental projects (Abderrezzak *et al.* 2015). After being discharged into a river, contaminants and effluents mix with water of the river being transported to the downstream (Seo & Cheong 1998). The effluent is spread vertically, transversely, and longitudinally by advective and dispersive transport processes. In a shallow stream, after contamination is rapidly mixed throughout the depth, the transmission will occur in the longitudinal and transverse directions (Ahmad *et al.* 2011). A full cross-sectional mix will not be achieved, unless the pollutant travels the long distances which are generally not within the length of practical interest (Beltaos 1980). The length required for full cross-sectional mixing of contaminations is approximately 20 and 200 times the upper width for a rough and a smooth flow, respectively (Fischer 1967). Transverse mixing plays an important role in determining the effect of contaminants under steady-state conditions. This parameter has an important effect in water quality management; especially in a case of point source discharges or tributary inflows (Rutherford 1994; Boxall & Guymer 2003). According to Figure 1, for the effluent mixing process in rivers, three stages are considered: (1) mixing near to the discharging point due to initial momentum and flow buoyancy (between A and B zones); (2) transverse mixing due to turbulence

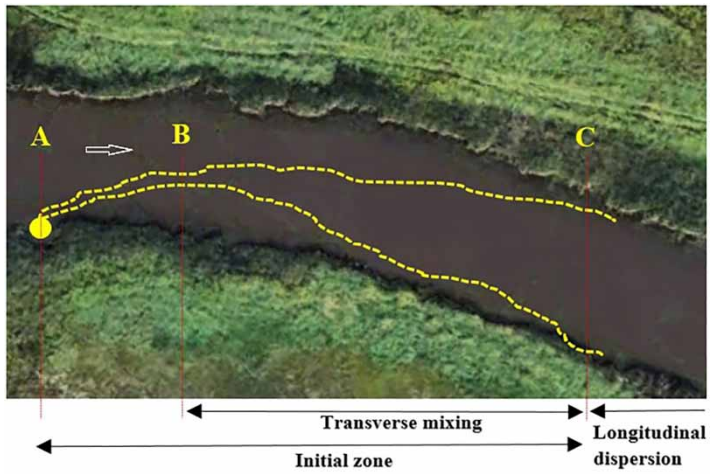

Figure 1 | General steps of pollution dispersion in a stream (Fischer *et al.* 1979).

(secondary turbulence transfer) and its secondary flows (between B and C zones); and (3) dispersion due to longitudinal shear flow (after C zone) (Fischer *et al.* 1979).

The distribution of tracer concentration can be written in a two-dimensional model according to the principle of mass conservation (Rutherford 1994; Sharma & Ahmad 2014):

$$\frac{\partial}{\partial t}(HC)+\frac{\partial}{\partial x}(HU_xC)+\frac{\partial}{\partial z}(HU_zC)=\frac{\partial}{\partial x}\left(H\varepsilon_x\frac{\partial C}{\partial x}\right)+\frac{\partial}{\partial z}\left(H\varepsilon_z\frac{\partial C}{\partial z}\right) \tag{1}$$

where t is the time; H is depth of flow (m); C is the depth-averaged tracer concentration (kg/m^3); z and x are the transverse and longitudinal directions, respectively; U_z and U_x are the velocities in the z and x directions (m/s), respectively; ε_z and ε_x are the depth-averaged dispersion coefficients in transverse and longitudinal directions (m^2/s). By assuming that longitudinal dispersion of tracer has not begun yet for the uniformly flowing stream, the time differentiation of Equation (1) will be zero (Sharma & Ahmad 2014). Also, by assuming a uniform flow and $U_z = 0$, Equation (1) can be simplified to:

$$U_x\frac{\partial C}{\partial x}=\varepsilon_z\frac{\partial^2 C}{\partial z^2} \tag{2}$$

The above equation has been used in many studies (Krishnappan & Lau 1977; Lau & Krishnappan 1981; Demetracopoulos 1994; Ahmad 2008; Aghababaei *et al.* 2017; Huai *et al.* 2018; Zahiri & Nezaratian 2020). More investigations on the role of the effective parameters in transverse mixing would be required due to the complexity of the transverse mixing mechanism (Aghababaei *et al.* 2017). Thus, predicting the transverse mixing coefficient (TMC) for known flow conditions in a stream for accounting the pollutant concentration at any location downstream of the injection site is genuinely essential (Azamathulla & Ahmad 2012). Generally, there are three approaches for predicting the TMC in stream mixing. Empirical methods have developed equations using the hydraulic and geometric dataset of rivers and experimental studies in order to establish a relationship for ε_z and theoretical methods have used the concept of shear flow to derive the dispersion coefficient (Baek & Seo 2013). Moreover, many researchers have recently used powerful predictive tools to find solutions for complex engineering problems. The significance of dispersion coefficients in water quality modeling and the complexity of the pollutant emission and mixing process have considerably increased the importance of using these tools (Zahiri & Nezaratian 2020). Soft computing techniques such as fuzzy-neural inference system-based principal component analysis (ANFIS-based PCA), particle swarm optimization method (PSO), artificial neural network (ANN), genetic expression programming (GEP), differential evolution (DE), decision tree (M5), support vector machine (SVM), and fuzzy-neural inference system (ANFIS) have been widely used to estimate longitudinal dispersion coefficient in streams by Parsaei *et al.* (2018), Alizadeh *et al.* (2017), Antonopoulos *et al.* (2015), Sattar & Gharabaghi (2015), Li *et al.* (2013), Etemad-Shahidi & Taghipour (2012), Azamathulla & Wu (2011) and Riahi-Madvar *et al.* (2009). Azamathulla & Ghani (2011), Azamathulla & Ahmad (2012), Aghababaei *et al.* (2017), and Zahiri & Nezaratian (2020), tried

to predict the TMC accurately by using decision tree (M5), multivariate adaptive regression splines (MARS), particle swarm optimization method (PSO), multiple linear regression (MLR), genetic algorithm (GA), genetic programming for symbolic regression (GPSR), and GEP. Soft computing techniques used by the above-mentioned researchers have less statistical errors and higher accuracy than empirical methods in TMC prediction (Zahiri & Nezaratian 2020). According to previous studies, there is a strong relationship between the TMC and channel parameters such as channel width, flow depth, shear velocity, friction factor, curvature and sinuosity (Fischer 1967; Beltaos 1979; Lau & Krishnappan 1981; Stefanovic & Stefan 2001; Boxall & Guymer 2003). Table 1 shows some of the most well-known equations proposed for calculating the TMC.

Each of these mentioned algorithms has its strengths and weaknesses that may not be able to predict complex phenomena such as TMC accurately. Selecting several meta-heuristic algorithms correctly and using them simultaneously will increase accuracy and decrease errors in target values' estimation. Selecting an algorithm as the main algorithm along with an auxiliary algorithm that can improve the weaknesses of the main algorithm will lead to developing a hybrid algorithm with higher performance. In previous investigations, several hybrid algorithms were used to estimate some of the complex phenomena and, consequently, the ability of these algorithms was proven completely (Pourbasheer *et al.* 2009; Wang *et al.* 2013; Li & Kong 2014; Zhou *et al.* 2016). In this study, two common algorithms were used to develop a hybrid algorithm: support vector machine (SVM) as the main algorithm and genetic algorithm (GA) as the auxiliary algorithm. Connecting GA to SVM allows us to estimate optimal values of SVM's adjustable parameters in the shortest time and increase predicting accuracy. The purpose of this study is developing an SVM-GA algorithm by using 232 published datasets and making a comparison of its performance with previous models. In addition, sensitivity analysis has been performed on the developed model to determine the effect of input parameters in the TMC modeling.

2. MATERIALS AND METHODS

2.1. Data

In the present study, 232 data points (see Supplementary material) were collected from the technical literature (Yotsukura *et al.* 1970; Holley & Abraham 1973; Krishnappan & Lau 1977; Beltaos 1979; Rutherford 1994; Jeon *et al.* 2007; Baek & Seo 2008;

Table 1 | Some of the empirical and data-driven models for estimation of TMC

Reference	Formula
Fischer & Park (1967)	$\varepsilon_z = 0.23HU_*$
Yotsukura *et al.* (1970)	$\varepsilon_z = 0.7HU_*$
Chau (2000)	$\varepsilon_z = 0.18HU_*$
Ahmad (2007)	$\varepsilon_z = 0.15HU_*$
Jeon *et al.* (2007)	$\varepsilon_z = 0.0291\left(\frac{W}{H}\right)^{0.463}\left(\frac{U}{U_*}\right)^{0.299}(S_n)^{0.733}(HU_*)$
Azamathulla & Ahmad (2012)	$\varepsilon_z = \left[\frac{\log(\log(\hat{U}+\hat{W}))}{\sin(-5.84\hat{W})-\hat{U}+2.765}+\frac{\hat{U}^{(\sin 8.22\hat{W})}}{\hat{W}^2-645.42}+\frac{0.35\hat{W}}{\ln(\hat{U})+2\hat{W}}\right]HU_*\ \hat{U}=\frac{U}{U_*}\text{ and }\hat{W}=\frac{W}{H}$
Aghababaei *et al.* (2017) (GPSR method)	$\varepsilon_z = 0.463+\left[0.464\left(\frac{U}{U_*}\right)\right]+\left[8.824\times10^{(-9)}\left(S_n\left(\frac{U}{U_*}\right)\right)\right]+\left[0.149\left(S_n\left[\left(\frac{U}{U_*}\right)+(2.306\times Fr\times S_n^2)-25.283\right]\right)\right]-\left[0.474\left(S_n\left[0.054\left(\frac{W}{H}\right)-20.371\right]\right)\right](HU_*)$
Zahiri & Nezaratian (2020) (M5 method)	$\varepsilon_z = 0.133\left(\frac{W}{H}\right)^{0.153}\left(\frac{U}{U_*}\right)^{-0.114}(Fr)^{-0.015}(S_n)^{0.168}HU_*$ $S_n \le 1.013$
	$\varepsilon_z = 0.236\left(\frac{W}{D}\right)^{0.044}(Fr)^{-0.045}(S_n)^{1.062}HU_*$ $S_n > 1.013$

ε_z is the TMC (m^2/s), H is the flow depth (m), U_* is a bed shear velocity (m/s), W is a channel width (m), S_n is sinuosity coefficient and Fr is a Froude number.

Lee & Seo 2013). It must be added that 183 and 49 dataset have been collected from straight and meandering streams, respectively. In addition, the dataset contains geometrical and hydraulic characteristics, including channel width, channel depth, average velocity, shear velocity, Froude number, sinuosity, and TMC. Sinuosity was used to demonstrate horizontal irregularities in meandering streams (Aghababaei *et al.* 2017). Table 2 illustrates a statistical analysis of all variables.

Table 2 implies that the studied cases varied from narrow rivers ($W/H< 10$) to very wide rivers ($W/H> 100$). $U/U*$, which is known as friction term and represents the hydrodynamic and roughness of the canal bed (Seo & Cheong 1998), varied from 0.026 to 28.571. This range of variations indicates the usage of a wide range of streams with various geometrical and hydraulic features in this study, the results of which can be related to many streams with different characteristics. The dataset was randomly divided into two sets, training (75% of the data) and testing (25% of the data). Although many unknown parameters may affect the TMC, according to previous studies, the key parameters affecting the mixing process during steady flow in natural streams can be stated as follows:

$$\varepsilon_z = f_1(U, U_*, W, H, \rho, \mu, S_f, S_n, g) \tag{3}$$

where ρ is the fluid density; μ is fluid viscosity; S_f and S_n are bed shape factor and sinuosity, respectively; and g is gravity. Fischer *et al.* (1979) and Jeon *et al.* (2007) expressed the relation below in terms of dimensionless parameters by using Buckingham Pi theorem:

$$\frac{\varepsilon_z}{HU_*} = f_2\left(\frac{U}{U_*}, \frac{W}{H}, \frac{U}{\sqrt{gH}}, \rho\frac{HU}{\mu}, S_f, S_n\right) \tag{4}$$

where U/U_* is the friction term; W/H is the channel width to flow depth ratio; $U/\sqrt{gH}$ is Froude number; and $\rho HU/\mu$ is Reynolds number. Bed shape factor, S_f, and sinuosity, S_n, indicate vertical and transverse irregularities in natural streams, respectively (Etemad-Shahidi & Taghipour 2012). By developing secondary currents and shear flow, transverse and vertical irregularities affect the mixing processes in streams (Seo & Cheong 1998). Generally, the flow in natural streams is usually fully turbulent, so Reynolds number could be eliminated from Equation (4) as a first approximation (Seo & Cheong 1998; Kashefipour & Falconer 2002). Bed shape factor S_f could also be eliminated from this equation as Froude number (Fr) and dimensionless roughness factor U/U_* can reflect the other effects of bed material roughness and bed slope (Sattar & Gharabaghi 2015). Finally, the best dimensionless form of ε_z based on previous findings such as those of Yotsukura & Sayre (1976), Deng *et al.* (2001), Jeon *et al.* (2007), Azamathulla & Ahmad (2012), Aghababaei *et al.* (2017), and Zahiri & Nezaratian (2020) can be written as follows:

$$\frac{\varepsilon_z}{HU_*} = f\left(\frac{U}{U_*}, \frac{W}{H}, Fr, S_n\right) \tag{5}$$

where $\frac{\varepsilon_z}{HU_*}$ represents the dimensionless parameter of ε_z and it will be used as the target parameter in this research. The correlations between all input and output parameters are displayed in Figure 2.

Based on Figure 2, there is no considerable correlation between the input variables, thus the problems that could arise in analysis from exaggerating the strength of the relations between variables, would be eliminated (Sattar & Gharabaghi 2015).

Table 2 | Descriptive statistics for the TMC database

Parameter	W	H	U	U_*	W/H	U/U_*	Fr	S_n	ε_z/HU_*	ε_z
Min	0.200	0.013	0.040	0.005	1.670	2.051	0.018	1.000	0.054	0.000034
Max	320.000	5.250	1.750	0.163	287.500	28.571	0.971	3.330	2.400	0.215
Avg	15.950	0.304	0.308	0.026	26.710	12.976	0.285	1.108	0.238	0.007
SD	51.237	0.709	0.271	0.023	34.995	5.447	0.181	0.371	0.249	0.025
Skewness	4.246	4.506	2.947	2.379	3.797	0.196	0.866	4.974	4.510	5.246

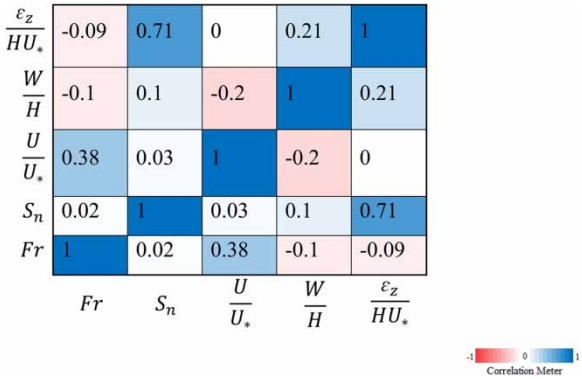

Figure 2 | Correlations between all input and output parameters.

It should be noted that the average of each parameter $\left(\frac{U}{U_*}, \frac{W}{H}, Fr, Sn, \frac{\varepsilon_z}{HU_*}\right)$ in training and testing subsets is equal to (13.36, 25.63, 0.29, 1.12, 0.25) and (11.91, 30.18, 0.27, 1.06, 0.20), respectively.

2.2. Support vector machine (SVM)

Vapnik (1995) proposed a nonlinear regression predicting method called support vector machine (SVM) which was usable to solve pattern recognition, highly nonlinear classification and regression problems. Maximizing the accuracy of prediction or minimizing the difference between the outputs and targets was the purpose of developing the SVM (Parsaie & Haghiabi 2017a, 2017b; Parsaie *et al.* 2019). For this purpose, the input parameters are mapped into a high-dimensional linear feature space by a nonlinear transformation to construct the optimal decision function. The dot product operation in the higher dimensional feature space is replaced by the kernel function in the original space, and by the finite sample training, the global optimal solution is obtained (Zhou *et al.* 2016). In the current study, SVM is used for predicting the TMC as the main algorithm, which is briefly described below.

If data $[(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)]$ is assumed as training set, where x_i is the input vector, $x \in R^n$, y_i is the output, $y \in R$ and n is the number of data pairs, the regression function of SVM which is called SVR will be formulated as follows:

$$y = f(x) = \omega^T \cdot \emptyset(x) + b \tag{6}$$

where ω^T represents the transposed form of ω vector; b is a bias; and ω can be obtained through some restricted rules. This function can describe the observed output y with an error tolerance ε. $\emptyset(x)$ would be considered as a nonlinear transfer function mapping the input vectors into a high-dimensional feature space which, theoretically, even a simple linear regression will be able to overcome the complexity of nonlinear regression of the input space (He *et al.* 2014). The tolerated errors within the extent of the ε-tube, as well as the penalized losses when data concern the outside of the tube, are defined by Vapnik's ε-insensitive loss function as:

$$l_\varepsilon(y_i) = \begin{cases} 0 & \text{for} |y_i - [\omega^T \cdot \emptyset(x_i) + b]| < \varepsilon \\ |y_i - [\omega^T \cdot \emptyset(x_i) + b]| - \varepsilon, & \text{for} |y_i - [\omega^T \cdot \emptyset(x_i) + b]| \geq \varepsilon \end{cases} \tag{7}$$

After that, the SVM problem can be formulated as the optimization problem as below:

$$Minimize_{(\omega,b,\xi,\xi^*)}\ R : \frac{1}{2}\omega^T \cdot \omega + C\left(\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \tag{8}$$

$$Subject\ to \left\{ \begin{array}{c} y_i - \omega^T \cdot \emptyset(x_i) - b_i \leq \varepsilon + \xi_i \\ \omega^T \cdot \emptyset(x_i) + b_i - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{array} \right\} i = 1, 2, \ldots, n \tag{9}$$

where the constant C is called a penalty factor and $C > 0$ shows the penalty degree of the sample with error exceeding ε (Liu & Jiao 2011). Here, the value of C is set to 1 which shows the complexity of the model is as important as the empirical error. Also, ξ_i and ξ_i^* are introduced as slack variables that specify the upper and lower errors of training subject to the error tolerance ε. These variables express the distance difference between actual values and the corresponding boundary values of ε-tube. Figure 3 depicts the mentioned situation graphically. SVM reduces under-fitting and over-fitting problems by minimizing $\frac{1}{2}\omega^T \cdot \omega$ and $C(\sum_{i=1}^{n}(\xi_i + \xi_i^*))$ which are called the regularization and training error terms, respectively.

Thus, the dual Lagrangian form will be yielded as follows by considering Lagrangian multipliers and Karush–Kuhn–Tucher condition in Equation (9):

$$Maximize\ L(\alpha_i, \alpha_i^*):\ \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) \tag{10}$$

$$Subject\ to \begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \tag{11}$$

where α_i and α_i^* are Lagrangian multipliers that satisfy equalities; $\alpha_i \times \alpha_i^* = 0$, and also, $L(\alpha_i, \alpha_i^*)$ represents the Lagrange function. The Lagrange multiplier terms $(\alpha_i - \alpha_i^*)$ related to the data accumulating the inside of the ε-insensitive tube will be considered to be zero. The final regression function is calculated only by using the datasets with non-zero coefficients $(\alpha_i - \alpha_i^*)$ which are known as the support vectors. There are two groups of support vectors: margin support vectors and error support vectors (Noori *et al.* 2011). In the first group, the support vectors have absolute values of the weights $|\alpha_i - \alpha_i^*|$ less than C and in the second group, equal to C. In other words, the support vectors, which are located outside and on the margin of the insensitive tube, are called the error support vectors and the margin support vectors, respectively (Figure 3). For changing the dimensionality of the input space to perform the regression or classification task with more confidence, kernel functions are used (Azamathulla & Wu 2011). These functions yield the inner products in the feature space $\emptyset(x_i)$ and $\emptyset(x_j)$. A kernel function plays the most significant role to simplify the learning process by changing the representation of the data in the feature space. Thus, although the data may be non-separable in the original input space, an appropriate choice of a kernel function allows the data to be highly separable in the feature space (Patil *et al.* 2012). If there is no prior knowledge about data features, radial basis function (RBF) will be recommended as one of the most popular kernel functions which is being used in different scientific fields (Roushangar & Koosheh 2015). For this reason, in this study,

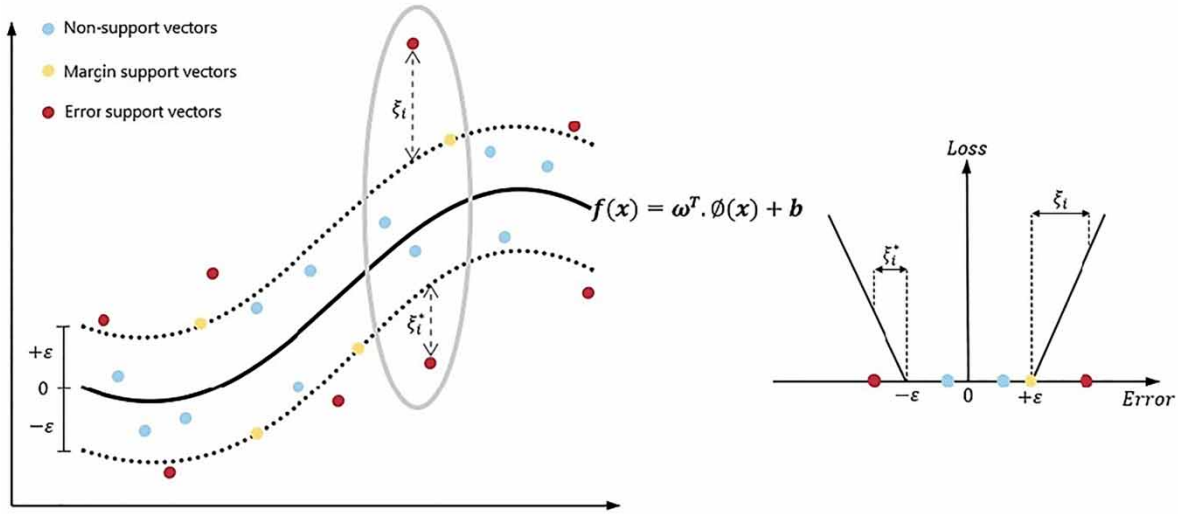

Figure 3 | Nonlinear SVM with Vapnik's e-insensitive loss function.

RBF was used as the kernel function of the SVM model for the TMC prediction.

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{12}$$

where $K(x_i, x_j) = \emptyset(x_i) \cdot \emptyset(x_j)$ is a kernel function and γ is the parameter of the RBF kernel function.

2.3. Genetic algorithm (GA)

According to the mechanisms of genetics and Darwin's natural selection principles, John Holland in 1975, proposed a heuristic search method and called it the genetic algorithm (GA). This method was named after biological processes of inheritance, mutation, natural selection, and the genetic crossover that happens when parents mate to produce offspring (Goldberg 1989). Technically, there are four differences between the structure of GA and other traditional optimization algorithms (Goldberg 1989):

- The GA typically uses a coding of the decision variable set instead of decision variable itself.
- The GA searches from a population of decision variable sets instead of a single decision variable set.
- The GA uses the objective function itself instead of the derivative information.
- The GA algorithm uses probabilistic instead of deterministic, search rules.

In the last decade, GA has successfully been used to solve some problems such as fitting nonlinear regression to data, optimizing simulation models, solving systems of nonlinear equations, and machine learning (Deb 1998). Generally, a GA has five major components to solve a particular problem that are briefly described below:

1 At the first, n chromosomes generate a population randomly that are known as candidate solutions to the problem.
2 A special fitness function evaluates the fitness of each chromosome. In the present study, efficiency coefficient (EC) was used as the fitness function and it can be written as:

$$EC = 1 - \frac{\sum_{i=1}^{N} (d_i - y_i)^2}{\sum_{i=1}^{N} (d_i - \bar{d})^2} \tag{13}$$

where N represents the total number of a testing data and y_i is the predicted value. d_i is the observed value and $\bar{d}$ is the mean of the observed values.

3 The following steps will be repeated until n offsprings have been created:
   (a) Selection: This operator selects the best chromosomes in pairs from the population to play the role of parents and reproduce two offspring. The more appropriate chromosomes have more chances to be selected.
   (b) Crossover: This operator randomly chooses a locus between a couple of chromosomes to form two offspring.
   (c) Mutation: This operator creates new chromosomes by flipping some of the bits in the chromosomes randomly.
4 Replace the current population with the new population.
5 If the stopping condition is satisfied, the best solution is returned in the current population, otherwise step 2 should be performed again.

The applied GA method settings in the present study are shown in Table 3.

Table 3 | Genetic algorithm settings

Population size	250
Number of generations	10
Elitism	12
Crossover probability	0.8
Mutation probability	0.1
Crossover function	Scatter
Mutation function	Gaussian

2.4. Genetic algorithm-based support vector machine

In this study, at first, the training data (input and target parameters) are presented to the GA-SVM algorithm. Then, GA randomly generates an initial population of unknown SVM's parameters (C, ε, and γ) to determine their optimal values to approach the best prediction with the lowest error and the highest accuracy. The fitness function examines the performance of each model. The secondary population of SVM's parameters is created by using the operators of GA (mutation, crossover, and selection) to obtain the optimal values of parameters and then these parameters are introduced to the SVM algorithm, again. This cycle is continued until the value of the fitness function is near or equal to the stopping conditions of the algorithm. Therefore, model outputs are expected to be closer to the target values at each cycle. In the GA-SVM algorithm, both algorithms operate separately but help each other in order to simplify the problem. In other words, first, SVM starts modeling by using the random parameters generated by GA, and GA continues the procedure of modeling until the optimal values of SVM's parameters are obtained. In this method, the GA algorithm tries to estimate the optimal combination of three parameters (C, ε, and γ) in each cycle. C is known as a regularization parameter that must control the trade-off between maximizing the margin and minimizing the training error. Low C values will place insufficient stress on fitting the training data and high values of C make the algorithm over-fit the training data (Noori *et al.* 2011). Nevertheless, according to Wang *et al.* (2003), it can be concluded that the prediction error is rarely influenced by C. γ denotes the optimal width of the kernel function, while RBF with large γ allows the support vector to have a strong impact over a larger area. The type of noise present in data determines the optimal value for ε, which is usually unknown. There is a practical consideration of the number of resulting support vectors, even if enough knowledge of the noise is available for selecting an optimal value for ε (Liu *et al.* 2006). In the GA-SVM hybrid algorithm, GA automatically starts finding the mentioned parameters of SVM and provides the optimal values, while determining the optimal values of parameters in the SVM algorithm was done by trial-and-error process. The cross-validation, which is an improved version of the grid search method, described by Hsu *et al.* (2010), was used to find these three parameters. In v-fold cross-validation, after the training set was divided into v subsets of equal size, one subset is tested sequentially by applying the classifier trained on the remaining $v-1$ subset. Therefore, each instance of the whole training set is estimated once so the cross-validation accuracy is the percentage of correctly classified data. The general flowchart of GA-SVM is illustrated in Figure 4.

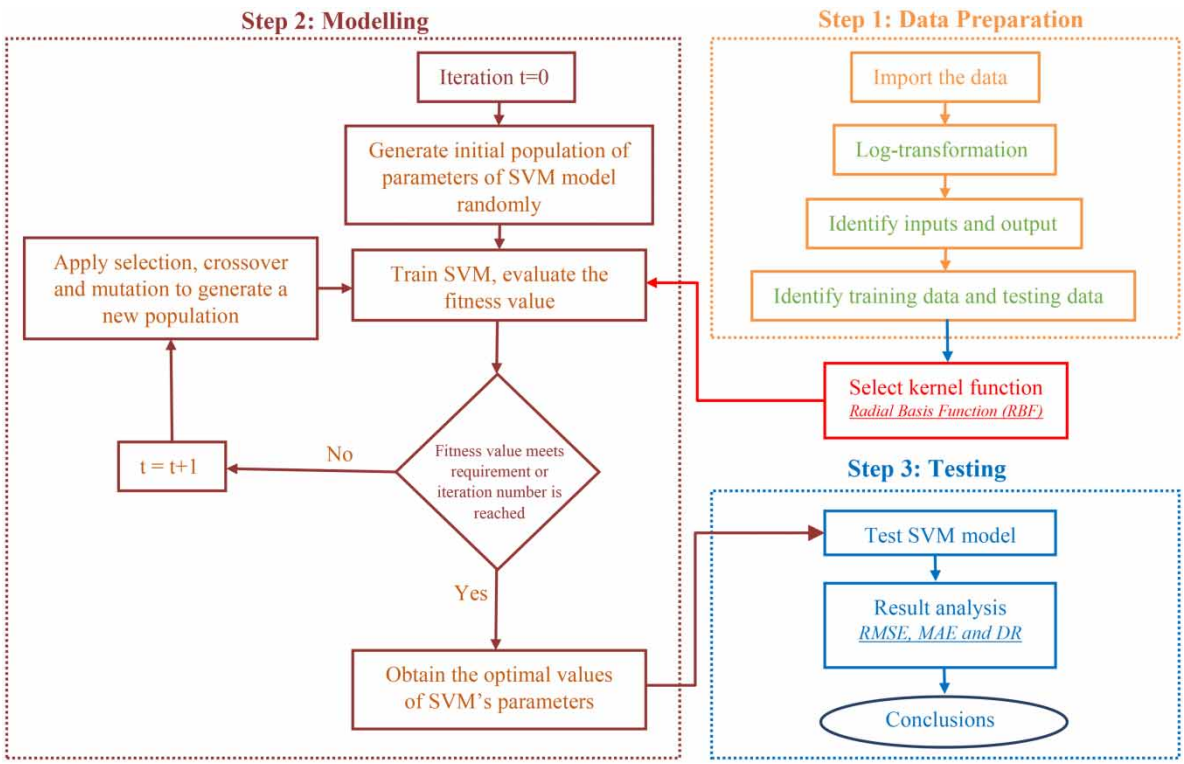

Figure 4 | General flowchart of GA-SVM algorithm.

In the present study, SVM and GA-SVM were applied by using RBF kernel function and input variables. Table 2 shows that all parameters used in this study have a right-skewed distribution. On the other hand, according to Figure 5, there is an abundance of outliers in the target and input parameters except *Fr* and U/U_*. Those observations which are uncommon and do not conform to the pattern of the majority of the data are called outliers (Rousseeuw & Van Zomeren 1990). The existence of outliers can cause increased error rates and reduce the accuracy of prediction. It can also lead to considerable distortions of statistic estimates when using either parametric or nonparametric tests (Zimmerman 1994, 1995, 1998). One of the simplest methods to tackle this problem is logarithmic transformations of parameters individually or collectively (Hubert & Van der Veeken 2008). Therefore, to reduce the negative effects of skewness and outliers on modeling, the whole dataset had been transformed into logarithmic scale and then the logarithmic parameters were used for modeling.

2.5. Model evaluation

In this study, both SVM and GA-SVM were used to estimate the TMC. The performances of these two models are assessed by evaluating the scatter plots between the observed and predicted results. In addition, the discrepancy ratio (DR), the root mean square error (RMSE), the mean of the absolute error (ME) and the accuracy were used as statistical parameters to evaluate the

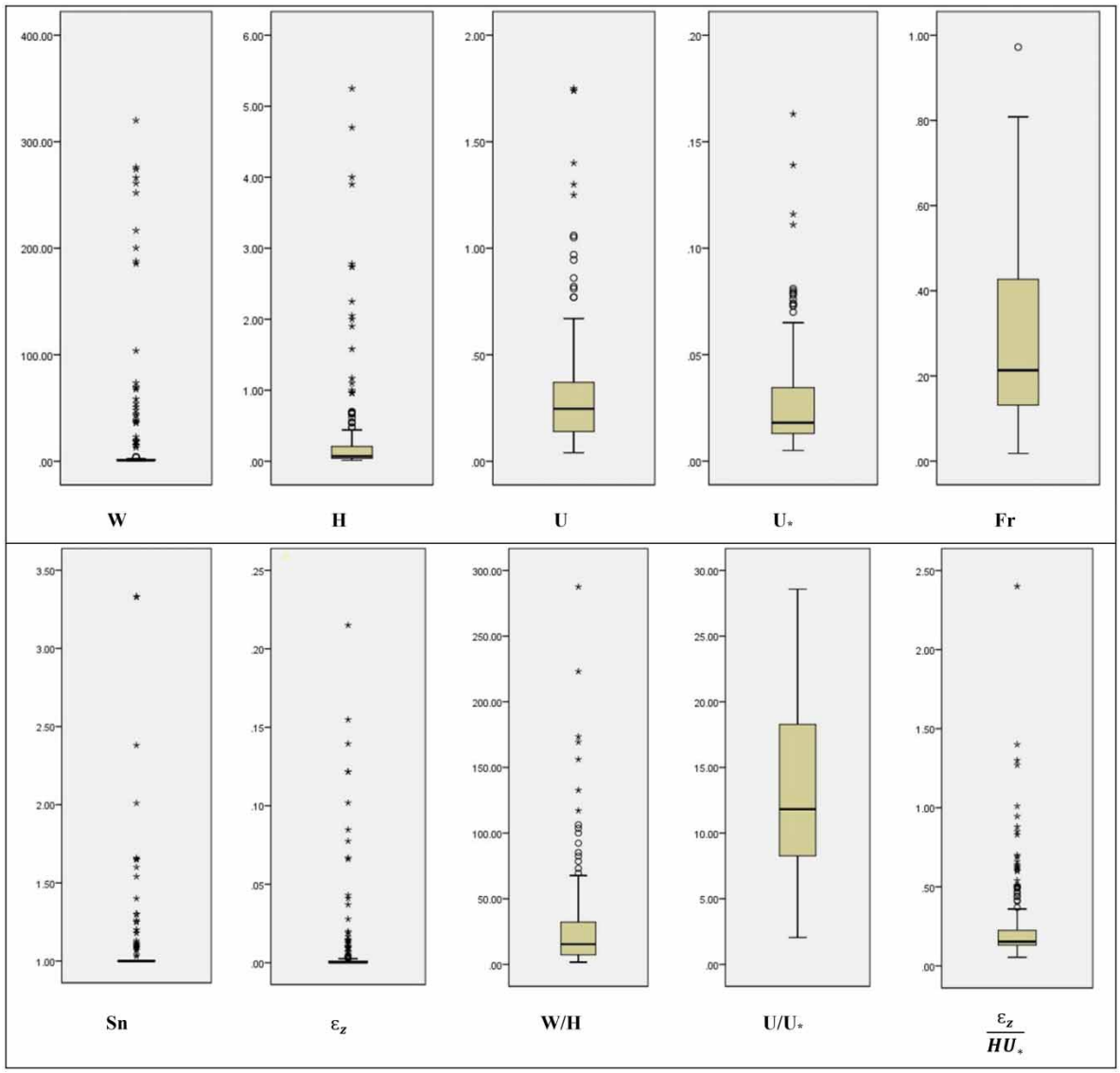

Figure 5 | Boxplots of all parameters with outliers (*).

performance of SVM, GA-SVM, and empirical models. Statistical indexes that were used in this study are expressed as:

$$\mathrm{DR} = \log\left(\frac{\varepsilon_z \mathrm{c}}{\varepsilon_z \mathrm{m}}\right) \tag{14}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{DR}_i)^2} \tag{15}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\mathrm{DR}_i| \tag{16}$$

where ε_zc and ε_zm are predicted and observed TMCs, respectively, and N is the total number of data points. If DR is equal to zero, there will be an exact match between the observed and predicted values. An overestimation (DR $>$ 0) or underestimation (DR $<$ 0) otherwise occurs. Previous researchers reported the percentage of DR values between -0.3 and 0.3 as an accuracy index (Seo & Cheong 1998; Kashefipour & Falconer 2002). In this research, in order to better evaluate the models' performance and accuracy, percentages of DR values between -0.15 and 0.15 were used as an accuracy index (Figure 6). As well, DR < -0.15 and DR > 0.15 have been considered as underestimation and overestimation beyond the precision range, respectively. A comparison of DR frequency could be used to determine the symmetry and skewness of TMC estimation by different models.

3. RESULTS AND DISCUSSION

For estimating TMC by using SVM, as was mentioned before, we first need to find the optimal values of three adjustable parameters of SVM (C, ε, and γ). During the grid search, all combinations of C, ε, and γ were tested for each cross-validation routine, where these parameters all ranged from 0 to 120. Finally, the optimum values of these three parameters were determined by using both GA and grid search algorithms. These values are presented in Table 4. According to Table 4, although both GA and grid search algorithms estimate parameter C to be approximately the same, their estimations were different for the other two parameters. It should be noted that GA does not estimate the optimal value of each parameter separately. This algorithm estimates only the optimal combination of the three parameters.

The performances of SVM, GA-SVM, and the previous methods in TMC estimation by using the mentioned statistical indexes are presented in Table 5.

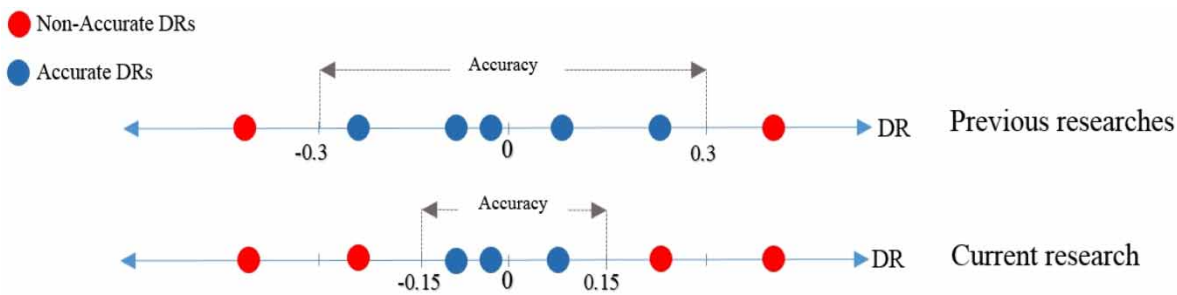

Figure 6 | Comparison of accuracy index between previous studies and the current study.

Table 4 | Optimal parameters of GA-SVM and SVM models

Models	Method	C	ε	γ
GA-SVM	GA	3.01	0.15	0.47
SVM	Grid Search	3.00	0.01	1.00

Table 5 | Performances of various methods on TMC estimation

Models	(DR < −0.15)	(−0.15 < DR < 0)	(0 < DR < 0.15)	(0.15 < DR)	Accuracy%	MAE	RMSE
Fischer & Park (1967)	15.086	9.052	19.397	56.466	28.448	0.228	0.270
Yotsukura *et al.* (1970)	2.155	1.724	6.466	89.655	8.190	0.588	0.626
Chau (2000)	19.397	12.931	52.586	15.086	65.517	0.180	0.255
Ahmad (2007)	25.431	28.017	41.810	4.741	69.828	0.169	0.273
Jeon *et al.* (2007)	12.931	13.362	31.034	42.672	44.397	0.188	0.233
Azamathulla & Ahmad (2012)	31.034	31.466	35.345	2.155	66.810	0.180	0.287
Aghababaei *et al.* (2017)	12.069	37.931	42.672	7.328	80.603	0.096	0.148
Zahiri & Nezaratian (2020)	11.638	31.466	44.397	12.500	75.862	0.113	0.149
GA-SVM (Train)	5.747	42.529	50.000	1.724	92.529	0.066	0.107
GA-SVM (Test)	10.345	32.759	50.000	6.897	82.759	0.097	0.139
SVM (Train)	5.747	42.529	48.851	2.874	91.379	0.044	0.096
SVM (Test)	12.069	32.759	48.276	6.897	81.034	0.097	0.152

Along with MAE, RMSE, and accuracy indexes, the balance between overestimation and underestimation values is also another important point in analyzing the models' performances. According to Table 5, among the previous regression models, the two models of Yotsukura *et al.* (1970) and Fischer & Park (1967), had the lowest performances in estimating the TMC with the accuracy of 8% and 28.5%, respectively. The two models of Aghababaei *et al.* (2017) and Zahiri & Nezaratian (2020) were able to have accurate performances in estimating TMC. The model of Aghababaei *et al.* (2017), based on GPSR method, with an accuracy of 80% and RMSE and MAE values of 0.148 and 0.096, respectively, and the simple data-driven-based model proposed by Zahiri & Nezaratian (2020) with a relatively good accuracy (75.8%) and the balance between overestimation and underestimation values were the most accurate regression-based models available to estimate this coefficient. Both GA-SVM and SVM algorithms had genuinely accurate and relatively similar performances. In the testing stage, both of them had the least error rates and the highest accuracy compared to the previous regression-based models. It should also be noted that although both models were based on the SVM algorithm, GA-SVM compared to SVM was able to improve the accuracy of the TMC estimation gently, in both training and testing stages by 1.15% and 1.7%, respectively. On the other hand, the grid search method is more time-consuming than GA, which make the GA-SVM model chosen for estimating TMC in this study. A comparison of the DR values of all expressions along with developed SVM and GA-SVM models is demonstrated in Figure 7. In addition, Figure 8 shows the performances of the developed SVM and GA-SVM in estimating the TMC for the two training and testing stages.

Based on Figure 7, the superiority of GA-SVM and SVM performance is obvious and both models have lower overestimation and underestimation values than the models of Aghababaei *et al.* (2017) and Zahiri & Nezaratian (2020). In addition, in Figure 8, the estimating accuracy by SVM and GA-SVM models are shown in training and testing stages, separately. The dataset used in this study included characteristics of straight and meandering streams. According to Table 6, the performance of both SVM and GA-SVM models in both straight and meandering streams was more accurate than the regression-based models. All models performed better in estimating the TMC in straight streams than meandering ones.

3.1. Sensitivity analysis

Sensitivity analysis helps researchers to determine which parameter has the most effect on reducing output uncertainty, and/or which parameters are negligible and can be eliminated from the final model (Nezaratian *et al.* 2018). In this study, a sensitivity analysis method was applied to determine the effect of each parameter on the performance of GA-SVM as the most accurate model in the TMC estimation. Five scenarios of the input parameter combinations were introduced to the GA-SVM algorithm for the TMC estimation. Table 7 presents the combination of inputs, absent parameters, SVM parameters, and the performance of each scenario in the testing stage, respectively.

As presented in Table 7, the effect of eliminating each input parameter on accuracy of final GA-SVM model was determined. In the table above, $\Delta_{\text{Accuracy\%}}$ expresses the difference between the final accuracy of each scenario and the overall

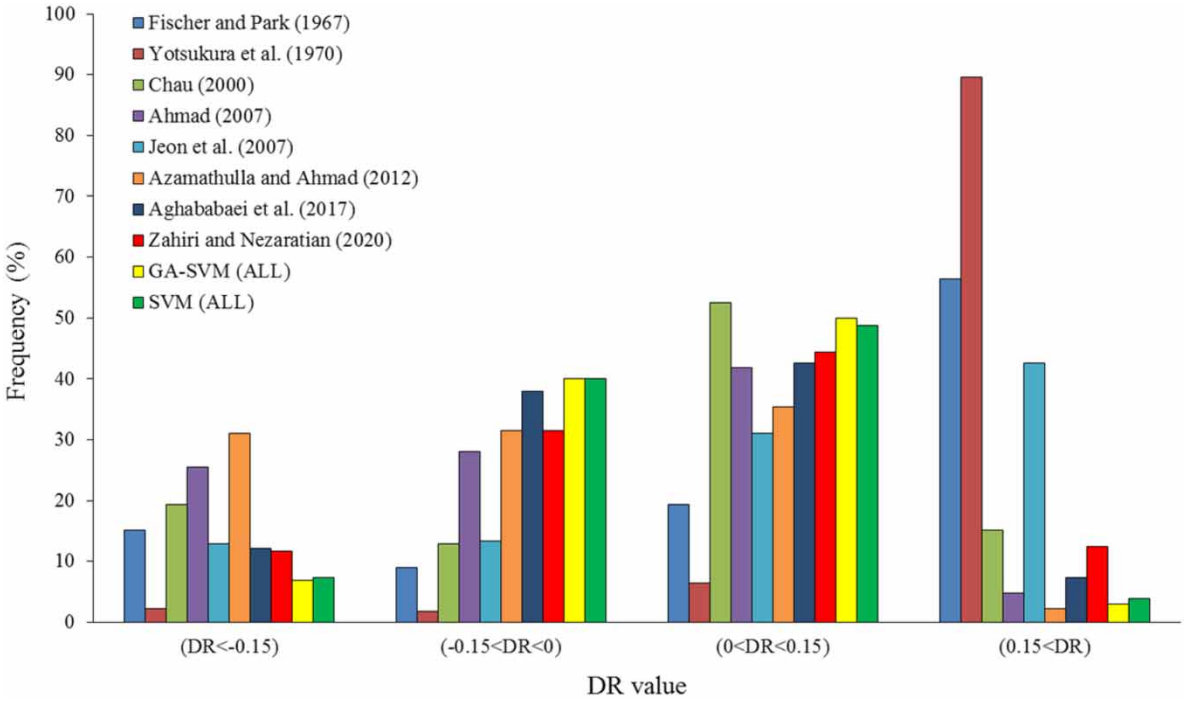

Figure 7 | Comparison of the DR values of different methods.

accuracy in the testing stage. It should be noticed that the above method significantly depends on the mathematical and theoretical structure of GA-SVM and may not be able to introduce the most effective parameter on TMC. However, analyzing Table 7 could help us, to some extent, on the effect of each input parameter on TMC estimation. The logic of input combination in scenario 5 was based on Figure 2. According to this figure, W/H and S_n have the highest correlation with the dimensionless parameter of the TMC while the lowest correlation belongs to U/U_* and Fr, respectively. Therefore, scenario 5 was used to measure the impact of removing the least correlated parameters on modeling TMC by GA-SVM. According to Table 7, in scenario 1, by eliminating W/H from the input parameters, the accuracy increases by 1.725%. However, in scenario 2, when W/H was replaced with U/U_* in the input variables, the accuracy was improved by 3.488%. In addition, using the same analysis and considering scenario 3, it can be deduced that Fr is the least effective parameter on TMC estimation by using the GA-SVM algorithm. According to scenario 4, it can also be concluded that Sn is a most efficient parameter in the process of modeling TMC. In scenario 5, only inputs which had a correlation coefficient above 0 were used, so U/U_* and Fr were eliminated from the process. The result showed that there was a significant improvement in the accuracy of the final model, which increase the modeling accuracy by 8.26%. Table 7 demonstrated that reducing the number of input variables with low correlation with the target improved the performance of the final GA-SVM model. Eliminating the low correlated input variables could decrease the complexity of the modeling process and increase the accuracy. This finding agreed with the results of Zahiri & Nezaratian (2020) and Jeon *et al.* (2007), which showed that Sn and W/H are the most influential parameters in estimating the TMC, respectively.

4. CONCLUSION

In this study, SVM and GA-SVM algorithms were developed to estimate the transverse mixing coefficient that plays an important role in modeling the pollutant release into streams. For this purpose, three statistical indexes (accuracy, RMSE, and MAE) were used to determine the performance of different models. The results showed the superiority of the proposed model compared to well-known regression-based models. The results also showed that the two models proposed by Aghababaei *et al.* (2017) and Zahiri & Nezaratian (2020) had the highest accuracy in estimating the TMC, respectively. Dividing the dataset into two groups (straight and meandering streams) showed that SVM and GA-SVM are still more reliable than the previous models. In this study, the grid search method was used to develop the SVM algorithm and was much more time-consuming than the GA algorithm. Therefore, the GA-SVM model was chosen as the best model to estimate the TMC in

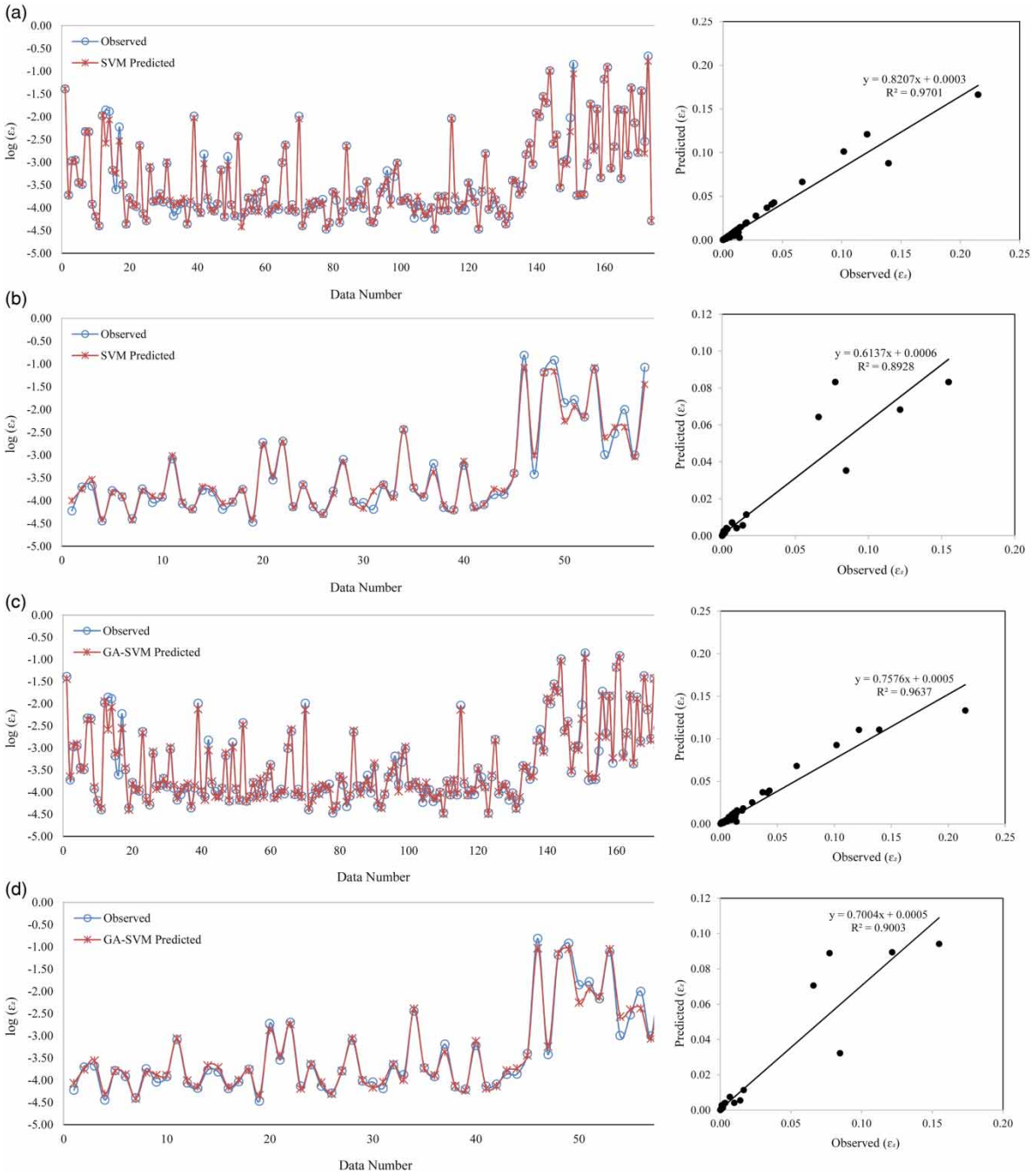

Figure 8 | The observed and predicted TMC (m^2/s) values by: (a) SVM in the training stage, (b) SVM in the testing stage, (c) GA-SVM in the training stage, and (d) GA-SVM in the testing stage.

Table 6 | Performances of various models using data of straight and meandering streams

	Straight			Meandering		
Models	**Accuracy%**	**MAE**	**RMSE**	**Accuracy%**	**MAE**	**RMSE**
Aghababaei *et al.* (2017)	85.246	0.082	0.124	63.265	0.150	0.216
Zahiri & Nezaratian (2020)	86.339	0.089	0.115	36.735	0.200	0.235
GA-SVM	93.443	0.063	0.099	77.551	0.113	0.164
SVM	91.803	0.049	0.098	77.551	0.083	0.155

Table 7 | Sensitivity analysis of GA-SVM scenarios

Scenario	Inputs	Absent	Parameters (C, ε, γ)	Accuracy%	MAE	RMSE	$\Delta_{\text{Accuracy\%}}$
1	U/U_*, Fr, S_n	W/H	7.75, 0.11, 0.30	84.483	0.064	0.110	1.725
2	W/H, Fr, S_n	U/U_*	5.47, 0.27, 0.20	86.207	0.074	0.131	3.448
3	W/H, U/U_*, S_n	Fr	4.38, 0.19, 0.25	89.655	0.062	0.117	6.896
4	W/H, U/U_*, Fr	S_n	2.33, 0.33, 1.59	81.034	0.087	0.124	−1.725
5	W/H, S_n	U/U_*, Fr	3.50, 0.47, 0.67	91.379	0.071	0.137	8.620

streams. Then, a sensitivity analysis was performed to determine the most effective input parameters in estimating the TMC by GA-SVM. Based on the sensitivity analysis, U/U_* and Fr had the least impact on GA-SVM performance in estimating TMC, and eliminating these two parameters improved the accuracy of the TMC estimation.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories (https://data.mendeley.com/datasets/2mm7jmp2g5/1).

REFERENCES

Abderrezzak, K. E. K., Ata, R. & Zaoui, F. 2015 One-dimensional numerical modelling of solute transport in streams: the role of longitudinal dispersion coefficient. *Journal of Hydrology* **527**, 978–989.

Aghababaei, M., Etemad-Shahidi, A., Jabbari, E. & Taghipour, M. 2017 Estimation of transverse mixing coefficient in straight and meandering streams. *Water Resources Management* **31** (12), 3809–3827.

Ahmad, Z. 2007 *Two-dimensional Mixing of Pollutants in Open Channels*. A technical report submitted to DST, New Delhi, India.

Ahmad, Z. 2008 Finite volume model for steady-state transverse mixing in streams. *Journal of Hydraulic Research* **46** (suppl. 1), 72–80.

Ahmad, Z., Azamathulla, H. M. & Zakaria, N. A. 2011 ANFIS-based approach for the estimation of transverse mixing coefficient. *Water Science and Technology* **63** (5), 1004–1009.

Alizadeh, M. J., Ahmadyar, D. & Afghantoloee, A. 2017 Improvement on the existing equations for predicting longitudinal dispersion coefficient. *Water Resources Management* **31** (6), 1777–1794.

Antonopoulos, V. Z., Georgiou, P. E. & Antonopoulos, Z. V. 2015 Dispersion coefficient prediction using empirical models and ANNs. *Environmental Processes* **2** (2), 379–394.

Azamathulla, H. M. & Ahmad, Z. 2012 Gene-expression programming for transverse mixing coefficient. *Journal of Hydrology* **434**, 142–148.

Azamathulla, H. & Ghani, A. 2011 Genetic programming for predicting longitudinal dispersion coefficients in streams. *Water Resources Management* **25** (6), 1537–1544.

Azamathulla, H. M. & Wu, F. C. 2011 Support vector machine approach for longitudinal dispersion coefficients in natural streams. *Applied Soft Computing* **11** (2), 2902–2905.

Baek, K. O. & Seo, I. W. 2008 Prediction of transverse dispersion coefficient using vertical profile of secondary flow in meandering channels. *KSCE Journal of Civil Engineering* **12** (6), 417–426.

Baek, K. O. & Seo, I. W. 2013 Empirical equation for transverse dispersion coefficient based on theoretical background in river bends. *Environmental Fluid Mechanics* **13** (5), 465–477.

Beltaos, S. 1979 Transverse mixing in natural streams. *Canadian Journal of Civil Engineering* **6** (4), 575–591.

Beltaos, S. 1980 Transverse mixing tests in natural streams. *Journal of the Hydraulics Division* **106** (10), 1607–1625.

Boxall, J. B. & Guymer, I. 2003 Analysis and prediction of transverse mixing coefficients in natural channels. *Journal of Hydraulic Engineering* **129** (2), 129–139.

Chau, K. W. 2000 Transverse mixing coefficient measurements in an open rectangular channel. *Advances in Environmental Research* **4** (4), 287–294.

Deb, K. 1998 Genetic algorithm in search and optimization: the technique and applications. In *Proceedings of International Workshop on Soft Computing and Intelligent Systems*. Machine Intelligence Unit, Indian Statistical Institute Calcutta, India, pp. 58–87.

Demetracopoulos, A. C. 1994 Computation of transverse mixing in streams. *Journal of Environmental Engineering* **120** (3), 699–706.

Deng, Z. Q., Singh, V. P. & Bengtsson, L. 2001 Longitudinal dispersion coefficient in straight rivers. *Journal of Hydraulic Engineering* **127** (11), 919–927.

Etemad-Shahidi, A. & Taghipour, M. 2012 Predicting longitudinal dispersion coefficient in natural streams using M5′ model tree. *Journal of Hydraulic Engineering* **138** (6), 542–554.

Fischer, H. B. 1967 The mechanics of dispersion in natural streams. *Journal of the Hydraulics Division* **93** (6), 187–216.

Fischer, H. B. & Park, M. 1967 *Transverse Mixing in A Sand-bed Channel*. US Geological Survey Professional Paper, 267–272.

Fischer, H. B., List, J. E., Koh, C. R., Imberger, J. & Brooks, N. H. 1979 *Mixing in Inland and Coastal Waters*. Academic Press, New York.
Goldberg, D. E. 1989 *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, New York, USA.
Haghiabi, A. H., Nasrolahi, A. H. & Parsaie, A. 2018 Water quality prediction using machine learning methods. *Water Quality Research Journal* **53** (1), 3–13.
He, Z., Wen, X., Liu, H. & Du, J. 2014 A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology* **509**, 379–386.
Holland, J. H. 1975 *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, USA.
Holley, E. R. & Abraham, G. 1973 Laboratory studies on transverse mixing in rivers. *Journal of Hydraulic Research* **11** (3), 219–253.
Hsu, C. C., Chen, M. C. & Chen, L. S. 2010 Intelligent ICA–SVM fault detector for non-Gaussian multivariate process monitoring. *Expert Systems with Applications* **37** (4), 3264–3273.
Huai, W., Shi, H., Yang, Z. & Zeng, Y. 2018 Estimating the transverse mixing coefficient in laboratory flumes and natural rivers. *Water, Air, & Soil Pollution* **229** (8), 252.
Hubert, M. & Van der Veeken, S. 2008 Outlier detection for skewed data. *Journal of Chemometrics: A Journal of the Chemometrics Society* **22** (3–4), 235–246.
Jeon, T. M., Baek, K. O. & Seo, I. W. 2007 Development of an empirical equation for the transverse dispersion coefficient in natural streams. *Environmental Fluid Mechanics* **7** (4), 317–329.
Kashefipour, S. M. & Falconer, R. A. 2002 Longitudinal dispersion coefficients in natural channels. *Water Research* **36** (6), 1596–1608.
Krishnappan, B. G. & Lau, Y. L. 1977 Transverse mixing in meandering channels with varying bottom topography. *Journal of Hydraulic Research* **15** (4), 351–370.
Lau, Y. L. & Krishnappan, B. G. 1981 Modeling transverse mixing in natural streams. *Journal of the Hydraulics Division* **107** (2), 209–226.
Lee, M. E. & Seo, I. W. 2013 Spatially variable dispersion coefficients in meandering channels. *Journal of Hydraulic Engineering* **139** (2), 141–153.
Li, X. Z. & Kong, J. M. 2014 Application of GA-SVM method with parameter optimization for landslide development prediction. *Natural Hazards and Earth System Sciences* **14** (3), 525.
Li, X., Liu, H. & Yin, M. 2013 Differential evolution for prediction of longitudinal dispersion coefficients in natural streams. *Water Resources Management* **27** (15), 5245–5260.
Liu, H. B. & Jiao, Y. B. 2011 Application of genetic algorithm-support vector machine (GA-SVM) for damage identification of bridge. *International Journal of Computational Intelligence and Applications* **10** (4), 383–397.
Liu, H., Yao, X., Zhang, R., Liu, M., Hu, Z. & Fan, B. 2006 The accurate QSPR models to predict the bioconcentration factors of nonionic organic compounds based on the heuristic method and support vector machine. *Chemosphere* **63** (5), 722–733.
Nezaratian, H., Zahiri, J. & Kashefipour, S. M. 2018 Sensitivity analysis of empirical and data-driven models on longitudinal dispersion coefficient in streams. *Environmental Processes* **5** (4), 833–858.
Noori, R., Karbassi, A. R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M. H., Farokhnia, A. & Gousheh, M. G. 2011 Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology* **401** (3–4), 177–189.
Parsaie, A. & Haghiabi, A. H. 2015 Predicting the longitudinal dispersion coefficient by radial basis function neural network. *Modeling Earth Systems and Environment* **1** (4), 1–8.
Parsaie, A. & Haghiabi, A. H. 2017a Mathematical expression of discharge capacity of compound open channels using MARS technique. *Journal of Earth System Science* **126** (2), 20.
Parsaie, A. & Haghiabi, A. H. 2017b Numerical routing of tracer concentrations in rivers with stagnant zones. *Water Science and Technology: Water Supply* **17** (3), 825–834.
Parsaie, A., Emamgholizadeh, S., Azamathulla, H. M. & Haghiabi, A. H. 2018 ANFIS-based PCA to predict the longitudinal dispersion coefficient in rivers. *International Journal of Hydrology Science and Technology* **8** (4), 410–424.
Parsaie, A., Haghiabi, A. H. & Moradinejad, A. 2019 Prediction of scour depth below river pipeline using support vector machine. *KSCE Journal of Civil Engineering* **23** (6), 2503–2513.
Patil, S. G., Mandal, S. & Hegde, A. V. 2012 Genetic algorithm based support vector machine regression in predicting wave transmission of horizontally interlaced multi-layer moored floating pipe breakwater. *Advances in Engineering Software* **45** (1), 203–212.
Pourbasheer, E., Riahi, S., Ganjali, M. R. & Norouzi, P. 2009 Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity. *European Journal of Medicinal Chemistry* **44** (12), 5023–5028.
Riahi-Madvar, H., Ayyoubzadeh, S. A., Khadangi, E. & Ebadzadeh, M. M. 2009 An expert system for predicting longitudinal dispersion coefficient in natural streams by using ANFIS. *Expert Systems with Applications* **36** (4), 8589–8596.
Roushangar, K. & Koosheh, A. 2015 Evaluation of GA-SVR method for modeling bed load transport in gravel-bed rivers. *Journal of Hydrology* **527**, 1142–1152.
Rousseeuw, P. J. & Van Zomeren, B. C. 1990 Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* **85** (411), 633–639.
Rutherford, J. C. 1994 *Longitudinal Dispersion. River Mixing*. John Wiley and Sons, Chichester, UK.
Sattar, A. M. & Gharabaghi, B. 2015 Gene expression models for prediction of longitudinal dispersion coefficient in streams. *Journal of Hydrology* **524**, 587–596.

Seo, I. W. & Cheong, T. S. 1998 Predicting longitudinal dispersion coefficient in natural streams. *Journal of Hydraulic Engineering* **124** (1), 25–32.

Sharma, H. & Ahmad, Z. 2014 Transverse mixing of pollutants in streams: a review. *Canadian Journal of Civil Engineering* **41** (5), 472–482.

Stefanovic, D. L. & Stefan, H. G. 2001 Accurate two-dimensional simulation of advective-diffusive-reactive transport. *Journal of Hydraulic Engineering* **127** (9), 728–737.

Vapnik, V. N. 1995 *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, USA.

Wang, W., Xu, Z., Lu, W. & Zhang, X. 2003 Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing* **55** (3–4), 643–663.

Wang, W. C., Xu, D. M., Chau, K. W. & Chen, S. 2013 Improved annual rainfall-runoff forecasting using PSO–SVM model based on EEMD. *Journal of Hydroinformatics* **15** (4), 1377–1390.

Yotsukura, N. & Sayre, W. W. 1976 Transverse mixing in natural channels. *Water Resources Research* **12** (4), 695–704.

Yotsukura, N., Fischer, H. B. & Sayre, W. W. 1970 *Measurement of Mixing Characteristics of the Missouri River Between Sioux City, Iowa, and Plattsmouth, Nebraska*. Water Supply Paper No. 1899-G. USGPO.

Zahiri, J. & Nezaratian, H. 2020 Estimation of transverse mixing coefficient in streams using M5, MARS, GA, and PSO approaches. *Environmental Science and Pollution Research* **27**, 14553–14566.

Zhou, C., Yin, K., Cao, Y. & Ahmed, B. 2016 Application of time series analysis and PSO–SVM model in predicting the bazimen landslide in the three gorges reservoir, China. *Engineering Geology* **204**, 108–120.

Zimmerman, D. W. 1994 A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology* **121** (4), 391–401.

Zimmerman, D. W. 1995 Increasing the power of nonparametric tests by detecting and downweighting outliers. *The Journal of Experimental Education* **64** (1), 71–78.

Zimmerman, D. W. 1998 Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education* **67** (1), 55–68.

First received 26 January 2021; accepted in revised form 30 May 2021. Available online 9 July 2021