Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers

Mohammad Rezaie Balf, Roohollah Noori, Ronny Berndtsson, Alireza Ghaemi and Behzad Ghiasi

ABSTRACT

The fate of pollutants in rivers is mainly affected by the longitudinal dispersion coefficient (K_x). Thus, improved K_x estimation could greatly enhance the water quality management of rivers. In this regard, evolutionary polynomial regression (EPR) was used to accurately predict K_x in rivers as a function of flow depth, channel width, and average and shear velocities. The predicted K_x by EPR modelling was compared with results obtained by more conventional K_x estimation formulas. Initial data analyses using general linear models of variance revealed that all input variables were statistically significant for K_x estimation. The calibrated EPR model showed good performance with coefficient of determination and root mean square error of 0.82 and 79 m^2/s, respectively. This is better that other more conventional estimation methods. Application of sensitivity analysis for the EPR model indicated that channel width, average velocity, shear velocity, and flow depth were the main variables in descending order that affected K_x variability. The introduced EPR estimation model for K_x can be incorporated in one-dimensional water quality models for improved simulation of solute concentration in natural rivers.

Key words | dispersion coefficient, evolutionary polynomial regression, pollution transport, rivers, sensitivity analysis

Mohammad Rezaie Balf
Alireza Ghaemi
Department of Civil Engineering,
Graduate University of Advanced Technology,
Kerman,
Iran

Roohollah Noori (corresponding author)
Behzad Ghiasi
Department of Environmental Engineering,
Graduate Faculty of Environment,
University of Tehran,
Enqelab Avenue, Qods Street, Azin Alley, P.O. Box: 14155-6135, Tehran,
Iran
E-mail: *noor@ut.ac.ir*

Ronny Berndtsson
Department of Water Resources Engineering & Center for Middle Eastern Studies,
Lund University,
Box 118, SE-221 00 Lund,
Sweden

INTRODUCTION

Rivers are some of the most important water resources exposed to pollution loads from natural and anthropogenic sources. Therefore, having a good knowledge on fate and transport of the pollutants is crucial to properly manage river water quality (Noori *et al.* 2012). The behaviour of a pollutant instantaneously introduced in rivers can be described by the three-dimensional advection-diffusion equation (3D-ADE) resulting from Fickian diffusion concepts. However, in the far downstream mixing zone, where variation concentrations in the vertical and horizontal directions are negligible, averaging the 3D-ADE over depth and width yields:

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

doi: 10.2166/aqua.2018.021

where C is cross-sectional averaged concentration, U is mean velocity at cross-section of river, t is time, x is longitudinal coordinate along the direction of mean flow, and K_x is longitudinal dispersion coefficient (Deng *et al.* 2002; Sahay & Dutta 2009).

The advection-dispersion equation has been widely used to investigate the behaviour of pollutants originating far upstream from non-steady point sources (Seo & Cheong 1998). When applying Equation (1) for simulation of solute concentration in rivers, K_x determination is important. The regular approach to determine K_x for rivers is tracer measurements. However, although tracer measurement approaches have been extensively performed for determination of K_x, their performance is confined by several limitations (Noori *et al.* 2017). For this reason, empirical methods for K_x

estimation have been widely applied by researchers. These methods include both user-friendly regression methods and more sophisticated black-box models based on artificial intelligence techniques (AIT) (Najafzadeh & Sattar 2015; Haghiabi 2016, 2017; Najafzadeh & Tafarojnoruz 2016; Alizadeh 2017; Alizadeh *et al.* 2017a; Parsaie & Haghiabi 2017a, 2017b; Wang *et al.* 2017). Since AIT are black-box models, their results are less applicable compared to more general techniques such as user-friendly regression methods for K_x estimation. In other words, it is not possible for other users to directly apply calibrated AIT to other cases of K_x estimation (i.e., different natural rivers) (Ebtehaj *et al.* 2015; Noori *et al.* 2015). Thus, there are problems involved in directly combining AIT with a physically based 1D model to enhance the accuracy of the spatiotemporal simulation of solute concentration in water bodies.

In this study, we suggest an alternative method for the prediction of K_x, evolutionary polynomial regression (EPR). The EPR is a hybrid regression method that was first introduced in the field of water resources by Giustolisi & Savic (2006). Abdul-Ghani *et al.* (2012) used EPR to predict the load of sediment in Malaysian rivers. Reyhani *et al.* (2013) investigated the EPR performance to compute the normalized flux, relative fouling and turbidity rejection in a wastewater treatment plant. EPR has also been applied in the field of water supply and river discharge prediction (Mounce *et al.* 2016; Rezaie-Balf & Kisi 2017).

Reported results from EPR applications have revealed its advantages, especially where there is a small subset of noisy or missing input data. In fact, EPR can be adapted to predict any arbitrary function in spite of missing input data. Meanwhile, due to the difficulty in collecting all relevant input data and the limited number of available datasets, K_x prediction does not include necessary input information (such as cross-mixing effect, transient storage, shearing advection and lateral mixing) that influence this parameter in rivers (Lanzoni *et al.* 2017). Excluding this information often results in major differences in predicted K_x as compared to measurements. Therefore, the EPR may be an appropriate technique for K_x estimation in natural rivers. In view of this, the paper first aims to present a prediction model for K_x based on the EPR method. Thereafter, the introduced EPR model performance is compared to some other more conventional methods for K_x estimation. Note that since the approach is a formula-based user-friendly method, researchers and practitioners can apply the developed method to predict K_x in any river type while applying their own datasets.

MATERIALS AND METHODS

K_x determination methods and data

In general, there are three methods for K_x determination in natural rivers: the flow structure method that directly quantifies the governing physical processes for dispersion; the method of moments based on measurements of a conservative pollutant concentration along the river; and empirical equations that are derived by correlation between K_x and geometric flow characteristics of the river. Due to the limitations of the first two methods (Wallis & Manson 2004), engineers tend to apply empirical approaches for K_x estimation. In this regard, the majority of developed empirical equations have supposed K_x as a function of geometry (including channel width B and flow depth H) and hydraulic conditions (including U and shear velocity U_*) such as Equation (2) (Seo & Cheong 1998; Kashefipour & Falconer 2002; Noori *et al.* 2011; Zeng & Huai 2014):

$$K_x = g(B,\ H, U, U_*) \tag{2}$$

Table 1 lists some of these empirical equations for K_x estimation. Therefore, the aim is to find a function g using the EPR method so that a best fit between observed and predicted K_x can be found. To apply EPR, a dataset including 149 patterns consisting of hydraulic and geometric characteristics measured in typical rivers was used (Supplementary material, Table S1, available with the online version of this paper).

There are some general guidelines in modelling to assure the modeller that the tuned model is comprehensive and can be used for other situations (new datasets). In our case, both calibration and validation patterns were selected so as to include high and low extreme values of K_x. In general, to have a representative model requires representative data. In a strict sense, this is only possible to check by using a variety of databases. However, Table 2 shows descriptive statistics of the datasets used in calibration and validation. The table shows that the largest K_x (1,486.5 m^2/s) is about twice the

Table 1 | Empirical equations for estimation of K_x

Author (s)	Equation
Elder (1959)	$K_x = 5.93U_*H$
Fischer (1975)	$K_x = 0.011((BU)^2/HU_*)$
Liu (1977)	$K_x = 0.18U_*H(B/H)^2(U/U_*)^2(U_*/U)^{1.5}$
Seo & Cheong (1998)	$K_x = 5.915U_*H(B/H)^{0.62}(U/U_*)^{1.428}$
Kashefipour & Falconer (2002)	$K_x = 10.612UH(U/U_*)$
Sahay & Dutta (2009)	$K_x = 2U_*H(B/H)^{0.96}(U/U_*)^{1.25}$
Etemad-Shahidi & Taghipour (2012)	$K_x = \begin{cases} 15.49U_*H(B/H)^{0.78}(U/U_*)^{0.11} & B/H \leq 30.6 \\ 14.12U_*H(B/H)^{0.61}(U/U_*)^{0.85} & B/H > 30.6 \end{cases}$
Li *et al.* (2013)	$K_x = 2.282U_*H(B/H)^{0.7613}(U/U_*)^{1.4713}$
Zeng & Huai (2014)	$K_x = 5.4(B/H)^{0.7}(U/U_*)^{0.13}HU$
Wang & Huai (2016)	$K_x = 17.648HU_*(B/H)^{0.3619}(U/U_*)^{1.16}$
Alizadeh *et al.* (2017b)	$K_x = \begin{cases} 5.319HU_*(B/H)^{1.206}(U/U_*)^{0.075} & (B/H) \leq 28 \\ 9.931HU_*(B/H)^{0.187}(U/U_*)^{1.802} & (B/H) \geq 28 \end{cases}$

Table 2 | Statistical indices of the parameters used for EPR model

Parameters	*B* (m)	*H* (m)	*U* (m/s)	U_* (m/s)	K_x (m²/s)
Training stage					
Maximum	196.6	8.2	1.73	0.33	1,486.5
Minimum	1.4	0.14	0.029	0.0016	0.2
Average	44.55	1.25	0.45	0.08	81.61
Validating stage					
Maximum	253.6	8.07	1.29	0.55	836.13
Minimum	10.97	0.32	0.08	0.0079	1.7
Average	60.83	1.55	0.51	0.09	87.04

size of the second largest K_x. A common way is to exclude such outliers (Tayfur & Singh 2005; Li *et al.* 2013; Disley *et al.* 2015). Although removing the largest K_x value may enhance K_x model performance, it limits the developed model application. Thus, this study aimed to develop a K_x estimation model using the EPR method while keeping outliers.

Development of EPR model

The EPR is a non-linear global stepwise regression method that provides mathematical structures on the basis of the evolutionary computing. In fact, genetic algorithm (GA) is applied in the general structure of EPR to define structure (e.g., number of mathematical terms, coefficients) of the mathematical expressions along with numerical regression (Najafzadeh *et al.* 2016; Bonakdari *et al.* 2017). Through the EPR model, the most common mathematical equation whose coefficients are optimized by means of GA, is written as (Laucelli & Giustolisi 2011):

$$y = \sum_{i=1}^{m} F(X, f(X),\ a_i) + a_0 \tag{3}$$

where y is the output, a_i are the parameters to be justified, F and f are functions that are calculated by the process and defined by the user, respectively, X is the input matrix, and m is the number of terms of the expression excluding a_0.

In the modelling process, the dataset is divided into the calibration and validation part. To select the best model, focusing on the model performance during the validation is important (Najafzadeh *et al.* 2017). Note that different factors such as number and range of inputs and type of selected functions (e.g., natural logarithmic, tangent hyperbolic, and exponential) play a key role in the EPR model performance (Najafzadeh *et al.* 2017). In EPR, the least squares approach is used to determine the setting parameters. The least square method assesses the setting parameters using the minimization of the sum of squared errors. However, having the calibration and validation dataset, a primary population of solutions is generated by application of GA and each parameter is known as the individual's chromosomes. The setting parameters are then assessed by minimizing the sum of squared errors. Finally, the general form of the EPR model is expressed as:

$$\hat{y} = a_0 + \sum_{i=1}^{m} a_i (X_1)^{ES(i,1)} \ldots (X_K)^{ES(i,K)} f\left((X_1)^{ES(i,K+1)}\right) \ldots f\left((X_K)^{ES(i,2K)}\right)$$
$$\hat{y} = a_0 + \sum_{i=1}^{m} a_i f\left((X_1)^{ES(i,1)} \ldots (X_K)^{ES(i,K)}\right)$$
$$\hat{y} = a_0 + \sum_{i=1}^{m} a_i (X_1)^{ES(i,1)} \ldots (X_K)^{ES(i,K)} f\left((X_1)^{ES(i,K+1)} \ldots (X_K)^{ES(i,2K)}\right)$$
$$\hat{y} = f\left(a_0 + \sum_{i=1}^{m} a_i (X_1)^{ES(i,1)} \ldots (X_K)^{ES(i,K)}\right) \quad (4)$$

where $\hat{y}$ is model predictions, K is number of independent predictor variables (inputs) and ES is user-defined function f matrix (coded as integers in the GA), whereas each element is a candidate used for each single input.

The inner functions of this model are considered to be linear, although they can be non-linear if the exponents are different from one. The initial validation of models produced by EPR is conducted on the basis of phenomena physical knowledge (Giustolisi & Savic 2006). Further details on EPR are described by Giustolisi & Savic (2006).

In this study, the software package EPR-MOGA-XL, working in MS-EXCEL environment, was used for application of EPR in the K_x estimation. The EPR model for prediction of K_x based on inputs B, H, U and U_* was used according to Equation (2). The setting parameters used to evaluate K_x estimation are given in Table 3. Figure 1 shows the different steps for developing the K_x estimation model using EPR.

It is worth noting that a sensitivity analysis was performed to determine the importance of each input for K_x estimation. In this regard, one parameter of Equation (2) was eliminated each time to evaluate its effect on the EPR performance.

Criteria for evaluation of model performance

In this section, to compare the EPR results with those obtained from more conventional regression methods, three indices were used: coefficient of determination (R^2), root mean square error ($RMSE$), and discrepancy ratio (DR) according to Seo & Cheong (1998):

$$R^2 = \left[\frac{\sum_{i=1}^{M}(K_{xi(Actual)} - \bar{K}_{x(Actual)})(K_{xi(Model)} - \overline{K}_{x(Model)})}{\sqrt{\sum_{i=1}^{M}(K_{xi(Actual)} - K_{x(Actual)})^2 . \sum_{i=1}^{M}(K_{xi(Model)} - K_{x(Model)})^2}}\right]^2 \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}(K_{(Model)} - K_{xi(Actual)})^2}{M}} \quad (6)$$

$$DR = \log\left(\frac{K_{xi}(Model)}{K_{xi}(Actual)}\right) \quad (7)$$

Table 3 | Details of setting parameters for development of EPR model

Description of parameter	Setting of parameters
Function set	Exponential
Type of model	Statistical
Type of presentation	$\hat{y} = a_0 + \sum_{i=1}^{m} a_i (X_1)^{ES(i,1)} \ldots (X_K)^{ES(i,K)} f\left((X_1)^{ES(i,K+1)}\right) \ldots f\left((X_K)^{ES(i,2K)}\right)$
Exponents range	[−2, −1.5, −1, −0.5, 0, 0.5, 1, 1.5, 2]
Number of mathematical terms	4
Bias (a_0) value	0

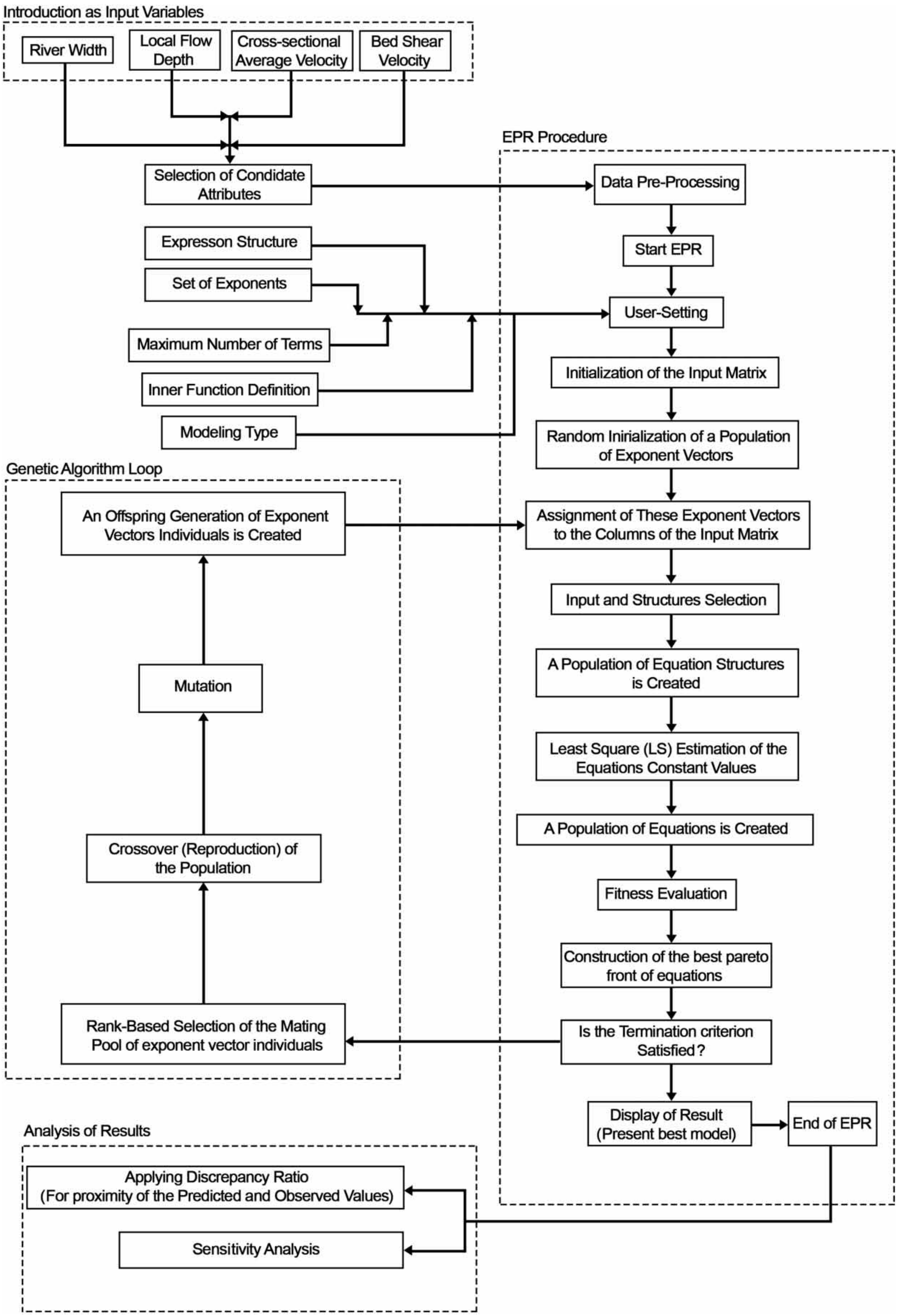

Figure 1 | Different steps for development of the K_x estimation model using EPR method.

where $K_{xi(Actual)}$ and $K_{xi(Model)}$ are the predicted (i.e., network output) and observed (i.e., target) values for the i^{th} K_x in the dataset, respectively. $\overline{K_{x(Actual)}}$ and $\overline{K_{x(Model)}}$ are the mean of $K_{xi(Actual)}$ and $K_{xi(Model)}$, respectively. M denotes the total number of observations.

For perfect fit, R^2 and $RMSE$ should equal one and zero, respectively. If DR is equal to zero, the model has a perfect fit. Otherwise, if DR is less (larger) than zero, the model overestimates (underestimates) K_x, respectively. The accuracy ranges of a model are defined as the percentage of DR that falls between −0.3 and 0.3 (Seo & Cheong 1998; Kashefipour & Falconer 2002).

RESULTS AND DISCUSSION

Model analysis

A general linear model analysis of variance (GLM-ANOVA) was applied to the results to statistically determine the significance of input variables. The GLM-ANOVA results revealed that the input variables B, H, U and U_* were statistically significant (p-value < 0.05) for the K_x estimation.

EPR results and comparison with other K_x estimation models

Different runs were performed by application of 103 datasets for calibrating the EPR model for K_x estimation; also, the rest of the data (46 datasets) kept away to validate the model performance.

After training of the EPR model, it provided several equations, as shown in Table 4. Based on the statistical indicators and a trade-off between accuracy and parsimony, Equation (9) had the best performance among all models. Note that GA population size, crossover probability rate and mutation probability rate for the best tuned model, i.e., Equation (9) were 40, 0.4 and 0.1, respectively.

The statistical indices show that Equation (9) could predict K_x with high accuracy (calibration). R^2 and $RMSE$ were equal to 0.80 and 90 m^2/s, respectively, which shows that calibration performed well.

Similar to calibration, validation showed that EPR produced a K_x with good accuracy ($R^2 = 0.82$, $RMSE = 79$ m^2/s). The EPR results for calibration and validation are summarized in Table 5. In addition, error plot resulting from calibration and validation of the chosen EPR model (Equation (9)) confirm these results (Figure 2).

Table 4 | List of explicit equations given by the EPR model

Model	Equation	R^2	$RMSE$ (m^2/s)
$K_x = +8.224\frac{U^{1.5}H^{0.5}B^{0.5}}{U_*^{0.5}}\ exp\ (-2U_*)$	(8)	0.76	93
$K_x = +9.1941\frac{U^2}{BU_*^2}\ exp\ (-1H + 2U - 2U_*) + 0.33128\frac{U^{1.5}HB}{U_*^{0.5}}\ exp\ (-0.5U_*)$	(9)	0.80	90
$K_x = +17.8066\frac{U^2H^{0.5}}{BU_*^2}\ exp\ (-1.5H + 2U - 2U_*) + 0.00018446\frac{U^2H^{0.5}B^2}{U_*^{0.5}}\ exp\ (+1.5U + 2U_*) + 40.5396U^{1.5}\ H\ B^{0.5}exp\ (-2U - 2U_*)$	(10)	0.79	90
$K_x = +17.8991\frac{U^2H^{0.5}}{BU_*^2}\ exp\ (-1.5H + 2U - 2U_*) + 0.006141\frac{U_*B^{1.5}}{U^2H^{1.5}}\ exp\ (-0.5H - 2U - 1.5U_*) + 37.6036U^{1.5}\ H\ B^{0.5}exp\ (-2U - 2U_*) + 0.0002361\frac{U^{1.5}H^{0.5}B^2}{U_*^{0.5}}exp\ (+1.5U + 2U_*)$	(11)	0.79	91

Table 5 | Results of K_x estimation models

Formula	R^2	RMSE (m²/s)
EPR, calibration stage	0.80	89
EPR, validation stage	0.82	79
Elder (1959)	0.07	186
Fischer (1975)	0.32	568
Liu (1977)	0.05	420
Seo & Cheong (1998)	0.76	92
Deng *et al.* (2001)	0.76	82
Kashefipour & Falconer (2002)	0.61	104
GA proposed by Sahay & Dutta (2009)	0.68	96
MT proposed by Etemad-Shahidi & Taghipour (2012)	0.52	144
DE proposed by Li *et al.* (2013)	0.74	85
Zeng & Huai (2014)	0.73	95
Wang & Huai (2016)	0.75	97
Alizadeh *et al.* (2017b)	0.73	86

Scatter plots of predicted and observed K_x for validation of the EPR model and the models developed based on evolutionary algorithms are shown in Figure 3. These models are differential evolution (DE) model (Li *et al.* 2013), M5' tree (MT) model (Etemad-Shahidi & Taghipour 2012), GA model (Sahay & Dutta 2009) and PSO model (Alizadeh *et al.* 2017b). Figure 3 shows that most predictions agree well with observations and that these fall on an almost straight line. However, a few predicted K_x values were greater than observations. The overall impression is, however, that the model yields satisfactory results.

The developed EPR performance was also compared with other K_x estimation models, such as the equations proposed by Elder (1959), Fischer (1975), Liu (1977), Seo & Cheong (1998), Deng *et al.* (2001), Kashefipour & Falconer (2002), Zeng & Huai (2014) and Wang & Huai (2016). This comparison is shown in Table 5. In addition, scatter plots of the equations for predicted and observed K_x in the validation step are shown in Figure 4. These results clearly show that the Elder (1959), Fischer (1975) and Liu (1977) equations performed least well for K_x estimation with R^2 equal to 0.07, 0.32 and 0.05, respectively. Table 5 and Figures 3 and 4 show that K_x estimation by the EPR model has the best R^2 and *RMSE* for validation.

The accuracy indicated by *DR* for the above-mentioned models are presented in Table 6 and Figure 5. According to these results, the accuracy for the EPR model (accuracy = 67.39%) is the best among all considered models. Kashefipour & Falconer (2002) and Alizadeh *et al.* (2017b) with accuracy equal to 56.51% are the second best models and the model proposed by Elder (1959) is found to be the least accurate model with accuracy equal to 2.17%.

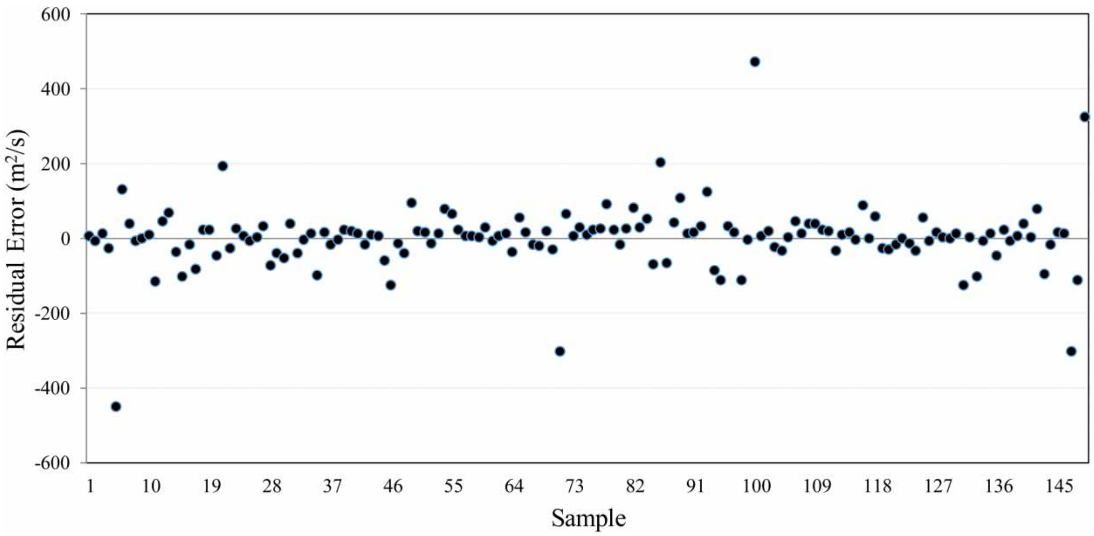

Figure 2 | Error plot using best-tuned EPR model during calibration and validation.

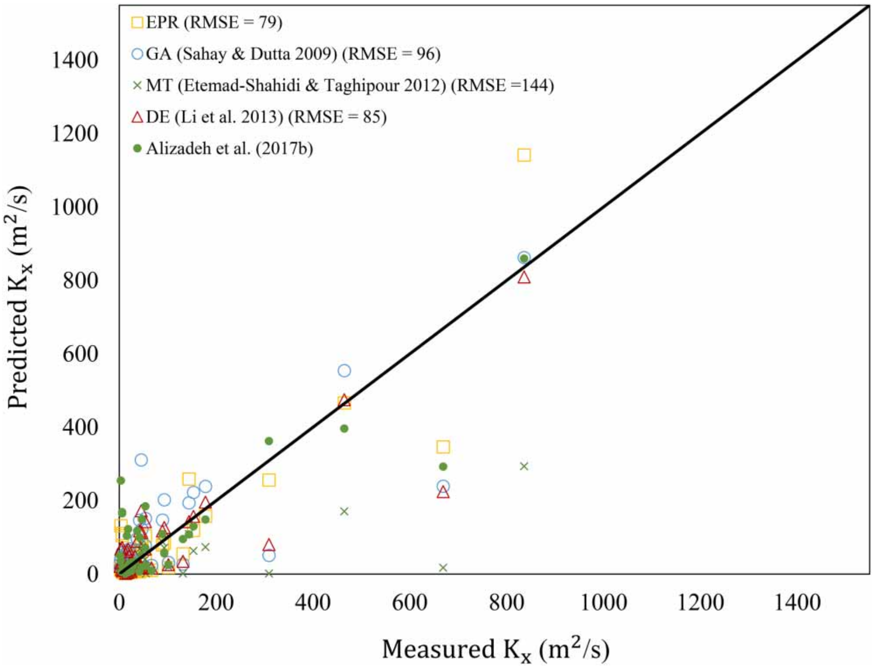

Figure 3 | Scatter plot of predicted and observed K_x for calibration and validation of the EPR model and the models developed based on evolutionary algorithms.

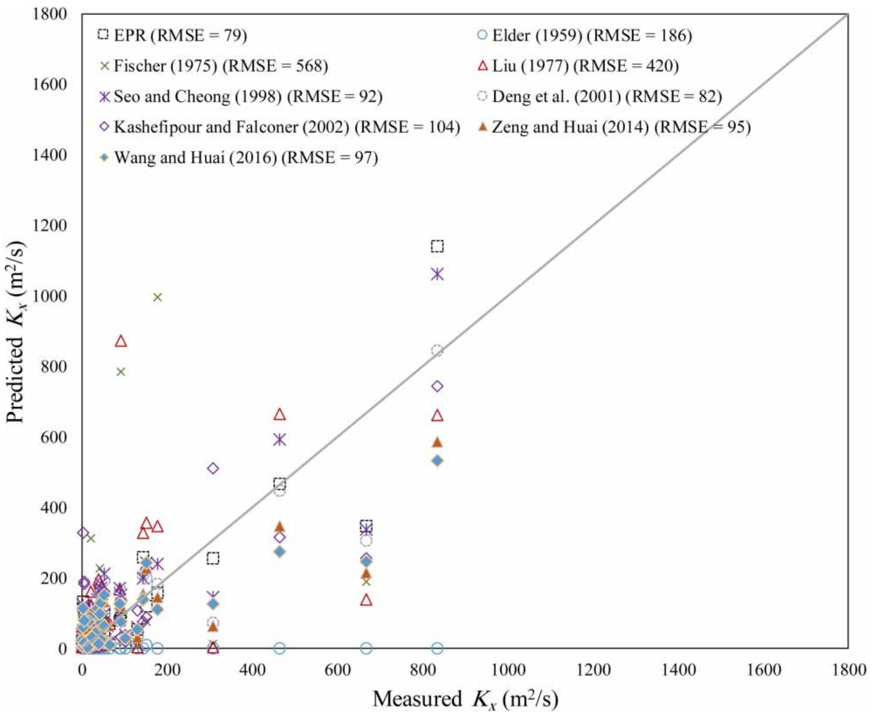

Figure 4 | Scatter plot of the EPR model and empirical equations for predicted and observed K_x in the validation step.

Table 6 | *DR* results for K_x estimation models

Techniques	$DR < -1$	$-1 < DR < -0.3$	$-0.3 < DR < 0$	$0 < DR < 0.3$	$0.3 < DR < 1$	$1 < DR$	Accuracy %
EPR	0	17.39	28.26	39.13	6.52	8.7	67.39
Elder (1959)	93.47	4.34	2.17	0	0	0	2.17
Fischer (1975)	8.69	19.56	4.34	23.91	36.95	6.55	28.25
Liu (1977)	4.34	15.21	8.69	36.95	28.26	6.55	45.64
Seo & Cheong (1998)	0	15.21	10.86	32.6	30.43	10.9	43.46
Deng *et al.* (2001)	0	19.56	13.04	26.08	32.6	8.72	39.12
Kashefipour & Falconer (2002)	6.52	15.21	43.47	13.04	13.04	8.72	56.51
GA (Sahay & Dutta 2009)	0	19.56	10.86	26.08	39.13	4.37	36.94
MT (Etemad-Shahidi & Taghipour 2012)	17.39	41.31	23.92	15.21	2.17	0	39.13
DE (Li *et al.* 2013)	0	10.86	10.86	30.43	19.56	28.29	41.29
Zeng & Huai (2014)	0	19.56	15.21	32.6	28.26	4.37	47.81
Wang & Huai (2016)		21.73	17.39	32.6	17.39	10.89	49.99
Alizadeh *et al.* (2017b)	0	19.56	21.73	34.78	13.04	10.89	56.51

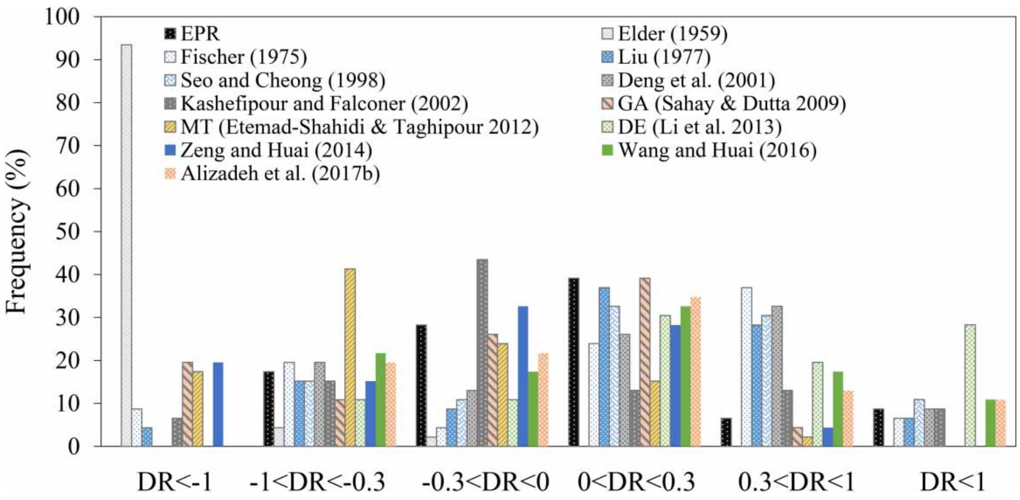

Figure 5 | *DR* accuracy for K_x estimation models.

Sensitivity analysis

The statistical error parameters from the sensitivity analysis are given in Table 7. These results demonstrate that B ($R^2 = 0.46$, $RMSE = 164$ m^2/s) was the most important variable in the K_x estimation whereas U_* ($R^2 = 0.78$, $RMSE = 86$ m^2/s) had the least influence. Other important variables for K_x estimation are U and H, respectively. Therefore, it is clear that B, U, U_* and H in descending order are the most important variables for K_x estimation. This is in line with the results of the MT model reported by Etemad-Shahidi & Taghipour (2012) and Noori *et al.* (2017).

Table 7 | Sensitivity analysis results

Input parameters	R^2	*RMSE* (m^2/s)
$K_x = g(H, B, U_*)$	0.74	93
$K_x = g(B, U, U_*)$	0.69	1,001
$K_x = g(U, H, U_*)$	0.46	164
$K_x = g(B, H, U)$	0.78	86

CONCLUSIONS

In this paper, the EPR technique was applied to predict K_x for natural rivers. In this regard, governing variables for the K_x, i.e., H, B, U and U_* were taken into account to develop the EPR model. The DE, MT and GA methods and regression equations proposed by other researchers were used in the comparison. Performance evaluation of K_x estimation models based on multiple error criteria (R^2 and *RMSE*) showed that the developed EPR model had the highest R^2 (0.82) and the least *RMSE* (79 m^2/s) for validation and outperformed other suggested models such as DE ($R^2 = 0.74$ and $RMSE = 85\ m^2/s$), Alizadeh *et al.* (2017b) ($R^2 = 0.73$ and $RMSE = 86$), GA ($R^2 = 0.68$ and $RMSE = 96\ m^2/s$) and MT ($R^2 = 0.52$ and $RMSE = 144\ m^2/s$). Additionally, results showed that the second best model was roughly the one proposed by Deng *et al.* (2001) with R^2 and *RMSE* equal to 0.75 and 82, respectively. The models proposed by Elder (1959), Fischer (1975) and Liu (1977) provided the worst performance with *RMSE* equal to 186, 568 and 420, respectively. Further investigations based on *DR* statistic revealed that the K_x estimation model developed by EPR and Elder (1959) had the best and the worst performance, respectively. In addition, based on SA results, B, U, U_* and H were the most important determining variables in descending order that affected K_x estimation for natural rivers, respectively.

Note, as we used data that have previously been used by other authors, and by comparing the results of these different models, we can say that we have results that are comparable to previously reported results.

REFERENCES

Abdul-Ghani, N. A., Shahin, M. A. & Nikoraz, H. R. 2012 Use of evolutionary polynomial regression (EPR) for prediction of total sediment load of Malaysian rivers. *Int. J. Eng.* **6**, 265–277.

Alizadeh, M. J. 2017 Discussion on 'Gene expression models for prediction of longitudinal dispersion coefficient in streams' by Sattar, A. M. A. & Gharabaghi, B. 2015. *J. Hydrol.* **524**, 587–596. https://doi.org/10.1016/j.jhydrol.2017.01.007.

Alizadeh, M. J., Shabani, A. & Kavianpour, M. R. 2017a Predicting longitudinal dispersion coefficient using ANN with metaheuristic training algorithms. *Int. J. Environ. Sci. Technol.* **14** (11), 2399–2410. https://doi.org/10.1007/s13762-017-1307-1.

Alizadeh, M. J., Ahmadyar, D. & Afghantoloee, A. 2017b Improvement on the existing equations for predicting longitudinal dispersion coefficient. *Water Resour. Manage.* **31** (6), 1777–1794. https://doi.org/10.1007/s11269-017-1611-z.

Bonakdari, H., Ebtehaj, I. & Akhbari, A. 2017 Multi-objective evolutionary polynomial regression-based prediction of energy consumption probing. *Water Sci. Technol.* **75** (12), 2791–2799. https://doi.org/10.2166/wst.2017.158.

Deng, Z. Q., Singh, V. P. & Bengtsson, L. 2001 Longitudinal dispersion coefficient in straight rivers. *J. Hydraul. Eng.* **127** (11), 919–927. https://doi.org/10.1061/(ASCE)0733-9429(2001)127:11(919).

Deng, Z., Bengtsson, L., Singh, V. P. & Adrian, D. D. 2002 Longitudinal dispersion coefficient in single-channel streams. *J. Hydraul. Eng.* **128** (10), 901–916. https://doi.org/10.1061/(ASCE)0733-9429(2002)128:10(901).

Disley, T., Gharabaghi, B., Mahboubi, A. & McBean, E. 2015 Predictive equation for longitudinal dispersion coefficient. *Hydrol. Process.* **29** (2), 161–172. https://doi.org/10.1002/hy0139p.1.

Ebtehaj, I., Bonakdari, H., Khoshbin, F. & Azimi, H. 2015 Pareto genetic design of group method of data handling type neural network for prediction discharge coefficient in rectangular side orifices. *Flow Meas. Instrum.* **41**, 67–74. https://doi.org/10.1016/j.flowmeasinst.2014.10.016.

Elder, J. W. 1959 The dispersion of a marked fluid in turbulent shear flow. *J. Fluid Mech.* **5** (4), 544–560. https://doi.org/10.1017/S0022112059000374.

Etemad-Shahidi, A. & Taghipour, M. 2012 Predicting longitudinal dispersion coefficient in natural streams using M5' model tree. *J. Hydraul. Eng.* **138** (6), 542–554. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000550.

Fischer, H. B. 1975 Discussion of 'simple method for predicting dispersion in streams' by R. S. McQuivey and T. N. Keefer. *J. Environ. Eng. Div.* **101** (3), 453–455.

Giustolisi, O. & Savic, D. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinform.* **8** (3), 207–222. https://doi.org/10.2166/hydro.2006.020.

Haghiabi, A. H. 2016 Prediction of longitudinal dispersion coefficient using multivariate adaptive regression splines. *J. Earth Syst. Sci.* **125** (5), 985–995. https://doi.org/10.1007/s12040-016-0708-8.

Haghiabi, A. H. 2017 Modeling river mixing mechanism using data driven model. *Water Resour. Manage.* **31** (3), 811–824. https://doi.org/10.1007/s11269-016-1475-7.

Kashefipour, S. M. & Falconer, R. A. 2002 Longitudinal dispersion coefficients in natural channels. *Water Res.* **36** (6), 1596–1608. https://doi.org/10.1016/S0043-1354(01)00351-7.

Lanzoni, S., Ferdousi, A. & Tambroni, N. 2017 River banks and channel axis curvature: effects on the longitudinal dispersion in alluvial rivers. *Adv. Water Resour.* **113**, 55–72. https://doi.org/10.1016/j.advwatres.2017.10.033.

Laucelli, D. & Giustolisi, O. 2011 Scour depth modelling by a multi-objective evolutionary paradigm. *Environ. Modell. Softw.* **26** (4), 498–509. https://doi.org/10.1016/j.envsoft.2010.10.013.

Li, X., Liu, H. & Yin, M. 2013 Differential evolution for prediction of longitudinal dispersion coefficients in natural streams. *Water Resour. Manage.* **27** (15), 5245–5260. https://doi.org/10.1007/s11269-013-0465-2.

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

Mounce, S. R., Blokker, E. J. M., Husband, S. P., Furnass, W. R., Schaap, P. G. & Boxall, J. B. 2016 Multivariate data mining for estimating the rate of discolouration material accumulation in drinking water distribution systems. *J. Hydroinform.* **18** (1), 96–114. https://doi.org/10.2166/hydro.2015.140.

Najafzadeh, M. & Sattar, A. A. 2015 Neuro-fuzzy GMDH approach to predict longitudinal dispersion in water networks. *Water Resour. Manage.* **29** (7), 2205–2219. https://doi.org/10.1007/s11269-015-0936-8.

Najafzadeh, M. & Tafarojnoruz, A. 2016 Evaluation of neuro-fuzzy GMDH-based particle swarm optimization to predict longitudinal dispersion coefficient in rivers. *Environ. Earth Sci.* **75**, 157. https://doi.org/10.1007/s12665-015-4877-6.

Najafzadeh, M., Balf, M. R. & Rashedi, E. 2016 Prediction of maximum scour depth around piers with debris accumulation using EPR, MT, and GEP models. *J. Hydroinform.* **18** (5), 867–884. https://doi.org/10.2166/hydro.2016.212.

Najafzadeh, M., Laucelli, D. B. & Zahiri, A. 2017 Application of model tree and evolutionary polynomial regression for evaluation of sediment transport in pipes. *KSCE J. Civ. Eng.* **21** (5), 1956–1963. https://doi.org/10.1007/s12205-016-1784-7.

Noori, R., Karbassi, A. R., Mehdizadeh, H., Vesali-Naseh, M. & Sabahi, M. S. 2011 A framework development for predicting the longitudinal dispersion coefficient in natural streams using an artificial neural network. *Environ. Prog. Sustain.* **30** (3), 439–449. https://doi.org/10.1002/e0478p.1.

Noori, R., Karbassi, A., Khakpour, A., Shahbazbegian, M., Badam, H. M. K. & Vesali-Naseh, M. 2012 Chemometric analysis of surface water quality data: case study of the Gorganrud River Basin, Iran. *Environ. Model. Assess.* **17** (4), 411–420. https://doi.org/10.1007/s10666-011-9302-2.

Noori, R., Deng, Z., Kiaghadi, A. & Kachoosangi, F. T. 2015 How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers? *J. Hydraul Eng.* **142** (1), 04015039. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001062.

Noori, R., Ghiasi, B., Shiekhian, H. & Adamowski, J. F. 2017 Estimation of the dispersion coefficient in natural rivers using a granular computing model. *J. Hydraul. Eng.* **143** (5), 04017001. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001276.

Parsaie, A. & Haghiabi, A. H. 2017a Computational modeling of pollution transmission in rivers. *Appl. Water Sci.* **7** (3), 1213–1222. https://doi.org/10.1007/s13201-015-0319-6.

Parsaie, A. & Haghiabi, A. H. 2017b Numerical routing of tracer concentrations in rivers with stagnant zones. *Water Sci. Technol.: Water Supply* **17** (3), 825–834. https://doi.org/10.2166/ws.2016.175.

Reyhani, A., Hemmati, M., Rekabdar, F. & Ahmadi, M. 2013 Application of evolutionary polynomial regression in ultrafiltration systems considering the effect of different parameters on oily wastewater treatment. *J. Petrol. Sci. Technol.* **3** (1), 9–19.

Rezaie-Balf, M. & Kisi, O. 2017 New formulation for forecasting streamflow: evolutionary polynomial regression vs. extreme learning machine. *Hydrol. Res.* nh2017283. https://doi.org/10.2166/nh.2017.283.

Sahay, R. R. & Dutta, S. 2009 Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm. *Hydrol. Res.* **40** (6), 544–552. https://doi.org/10.2166/nh.2009.014.

Seo, I. W. & Cheong, T. S. 1998 Predicting longitudinal dispersion coefficient in natural streams. *J. Hydraul. Eng.* **124** (1), 25–32. https://doi.org/10.1061/(ASCE)0733-9429(1998)124:1(25).

Tayfur, G. & Singh, V. P. 2005 Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. *J. Hydraul. Eng.* **131** (11), 991–1000. https://doi.org/10.1061/(ASCE)0733-9429(2005)131:11(991).

Wallis, S. G. & Manson, J. R. 2004 Methods for predicting dispersion coefficients in rivers. *P. I. Civil Eng.- Wat. M.* **157** (3), 131–141. https://doi.org/10.1680/wama.2004.157.3.131.

Wang, Y. & Huai, W. 2016 Estimating the longitudinal dispersion coefficient in straight natural rivers. *J. Hydraul. Eng.* **142** (11), 04016048. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001196.

Wang, Y. F., Huai, W. X. & Wang, W. J. 2017 Physically sound formula for longitudinal dispersion coefficients of natural rivers. *J. Hydrol.* **544**, 511–523. https://doi.org/10.1016/j.jhydrol.2016.11.058.

Zeng, Y. & Huai, W. 2014 Estimation of longitudinal dispersion coefficient in rivers. *J. Hydro-Environ. Res.* **8** (1), 2–8. https://doi.org/10.1016/j.jher.2013.02.005.

First received 31 January 2018; accepted in revised form 13 May 2018. Available online 6 June 2018