Estimation of natural streams longitudinal dispersion coefficient using hybrid evolutionary machine learning model

Leonardo Goliatta, Sadeq Oleiwi Sulaiman, Khaled Mohamed Khedher, Aitazaz Ahsan Farooque and Zaher Mundher Yaseen

Estimation of natural streams longitudinal dispersion coefficient using hybrid evolutionary machine learning model

Leonardo Goliatta, Sadeq Oleiwi Sulaiman, Khaled Mohamed Khedher, Aitazaz Ahsan Farooque and Zaher Mundher Yaseen

ABSTRACT
Among several indicators for river engineering sustainability, the longitudinal dispersion coefficient ($K_x$) is the main parameter that defines the transport of pollutants in natural streams. Accurate estimation of $K_x$ has been challenging for hydrologists due to the high stochasticity and non-linearity of this hydraulic-environmental parameter. This study presents a new hybrid machine learning (ML) model integrating a Gaussian Process Regression (GPR) and an evolutionary feature selection (FS) approach (i.e. Covariance Matrix Adaptation Evolution Strategy (CMAES)) to estimate $K_x$ in natural streams. The dataset consists of geometric and hydraulic river system parameters from 29 streams in the United States. The modeling results showed that the proposed model outperformed other models in the literature, producing more stable and accurate estimations. The FS approach evidenced the significance of the cross-sectional average flow velocity ($U$), channel width ($B$), and channel sinuosity ($\sigma$) to estimate the dispersion coefficient. In quantitative terms, the integrated GPR model with feature selection approach attained the minimum root mean square error (RMSE = 48.67) and maximum coefficient of determination ($R^2 = 0.95$). The proposed hybrid evolutionary ML model arises as robust, flexible and reliable alternative computer aid technology for predicting the longitudinal dispersion coefficient in natural streams.

ARTICLE HISTORY
Received 17 May 2021
Accepted 16 August 2021

KEYWORDS
Longitudinal dispersion coefficient; Gaussian processes regression; automated machine learning; natural streams

Introduction
Research background and $K_x$ empirical formulation

Various types of contaminants are discharged into water bodies across the world, both point and non-point sources, and these contaminants adversely affect the water quality (Noori et al., 2007; Tenebe et al., 2016). Hence, providing an accurate pollution dispersion analysis in rivers is a tedious issue for authorities responsible for monitoring and managing water resources. Even though pollution can disseminate in all directions of rivers (Desiante et al., 2021; Park & Hwang, 2019; Tayfur & Singh, 2005), the longitudinal dispersion usually is dominant at a considerable distance from the pollution source after discharge into the water body (Riahi-Madvar et al., 2009; Steele, 1991). As per several environmental and hydrological investigations, the most critical parameter when studying the longitudinal transport of pollution in rivers is the longitudinal dispersion coefficient ($K_x$) (Baek, 2019; Carter & Okubo, 1972; Dehghani et al., 2020; Etemad-Shahidi & Taghipour, 2012; Farzadkho et al., 2019, 2018; Hamidifar et al., 2015; Kashefpour & Falconer, 2002; Wang et al., 2017). Accurate estimation of $K_x$ is essential for numerous practical applications, such as environmental engineering, river engineering, intake designs, assessment of dangerous contaminants discharge into water bodies, etc. (M. J. Alizadeh, Ahmadyar et al., 2017). Several models exist which rely on the advection-dispersion equation shown in Equation (1); however, the modeling results are highly affected by the nature of the $K_x$ (Camacho Suarez et al., 2019).

$$K_x \frac{\partial^2 C}{\partial x^2} = \frac{\partial C}{\partial x} + \frac{\partial C}{\partial t}$$

where $K_x$ is the longitudinal dispersion coefficient, $t$ is the time, $x$ and $u$ are the longitudinal coordinate and mean cross-sectional velocity, respectively, and $C$...
is the averaged cross-sectional concentration (Noori et al., 2011). Regarding $K_x$, it can be determined experimentally, empirically, and theoretically (Camacho Suarez et al., 2019; Data, 2019; Deng et al., 2002, 2001; Disley et al., 2015; Milisić et al., 2019; Perucca et al., 2009; Seo & Baek, 2004; Seo & Cheong, 1998; Shin et al., 2019; Shucksmith et al., 2011; Swamee et al., 2000; Wang & Huai, 2016; Zeng & Huai, 2014). The direct experimental determination of the dispersion coefficient demands cost and time inefficient tracer studies and could be done only with rectangular flumes data (Etemad-Shahidi & Taghipour, 2012). For certain conditions, pipes $K_x$ can be formulated using the advection-diffusion equation subjected to specific water flow criteria (laminar or turbulent). However, for natural streams, the determination of the $K_x$ does not have a complete theoretical understanding owing to the stochasticity phenomena (Wang et al., 2017). In addition, there is no knowledge on the transverse profiles of both the flow depth and flow velocity (Deng et al., 2001; Paulson, 1970). The earlier predictive equations differ in their findings and are laden with certain levels of uncertainty and thus many research have focused on the development of more accurate empirical formulation (EF) for $K_x$ estimation (Altunkaynak, 2016; Nezaratian et al., 2018; Wang et al., 2017). Several studies have been devoted over the past decade to discovering the relationship between $K_x$ and the hydraulic geometric characteristics (Noori et al., 2021; Sahin, 2014; Shen et al., 2010; Tutmez & Yuceer, 2013); however, the presented EFs still have indicated their limitations on attaining high predictability performance.

$K_x$ estimation using artificial intelligence models: literature review

Over the past decades, the implementation of computer aid models reported promising results in modeling hydraulic engineering problems (Sharafati et al., 2021). The importance of artificial intelligence (AI) machine learning models in environmental and hydrological applications has increased over the years (Yaseen et al., 2019). One of the earliest studies conducted on the $K_x$ modeling was based on implementing an artificial neural network (ANN) model developed by Tayfur and Singh (2005) in rivers and natural streams. The evaluation of the model proved its capability in predicting the $K_x$ compared to the earlier suggested EFs. In another study by Tayfur (2006), the authors established fuzzy set, ANN, and regression-based models for the $K_x$ estimation in natural streams. The estimation results showed that the developed models outperformed the existing EFs as they are satisfactorily estimated the measured data with minimum errors. A fuzzy model was developed for $K_x$ estimation in natural channels by Fuat Toprak and Savci (2007). The study compared the performance of the developed fuzzy model with the measured data and the existing equations. The research results indicated that the feasibility of the fuzzy model outperformed the other models in terms of results reliability. Three ANN models based on different learning algorithms (i.e. the radial basis function neural network (RBFNN), feedforward backpropagation neural network (FBNN), and the generalized regression neural network) were tested for $K_x$ estimation (Toprak & Cigizoglu, 2008). The research findings showed that the developed model’s accuracy was higher than the accuracy of the other existing EFs. The capacity of support vector regression (SVR) and adaptive neuro-fuzzy inference system (ANFIS) models were inspected for $K_x$ estimation in natural streams by Noori et al. (2009). The study found that the SVR model performed better than the ANFIS model in achieving better threshold statistical analysis. The FBNN model with a 2D convergent flow tracer transport model was developed for improving the evaluation of transverse and longitudinal dispersivities from a convergent flow tracer test (Shieh et al., 2010). The developed model required less computational time and offered more accurate values of the transport parameter. The study also confirmed the effectiveness of the developed model for achieving fast and accurate transverse and longitudinal dispersivities evaluation for a field convergent flow tracer test. A genetic programming (GP) model was developed for modeling $K_x$ (Tu et al., 2015) and the study corroborated the capacity of the GP model in comparison with EFs. The use of the M5 model tree for $K_x$ estimation was conducted by Etemad-Shahidi and Taghipour (2012). The model was compared to the other existing equations in terms of performance based on error measures. The developed model performed better than the existing formulas and could be a valuable tool for $K_x$ estimation. Two gene expression programming (GEP) models were examined for $K_x$ estimation based on 150 published data sets of hydraulic and geometric parameters in natural streams (Sattar & Gharabagh, 2015). The analysis showed that the proposed relations were accurate, simple, and effective in $K_x$ estimation in natural streams.

Of the recent exploration for the advanced versions of AI models, hybrid version where two or more models are integrated for the purpose of improving the learning process has been noticed remarkably (Ghaemi et al., 2021; Tao, Al-Bedyry, et al., 2021; Tao, Habib, et al., 2021). The capacity of the particle swarm optimization (PSO) algorithm was explored to improve the performance of the neuro-fuzzy-based group method of data handling (NF-GMDH) model for $K_x$ estimation in rivers.
(Najafzadeh & Tafarojnoruz, 2016). The NF-GMDH-PSO model was compared with several AI models and traditional EFs in terms of accuracy performance. The proposed hybrid model evidenced its potential against the standalone AI models and the EFs. The reliability of ANN, ANFIS, and SVR models were studied for $K_x$ estimation in natural rivers (Noori et al., 2016). The study performed forward selection and gamma test to sort the input parameters in the order of their effects and relevance on $K_x$ estimation. The study’s outcome revealed less uncertainty in the SVR model compared to the ANN and ANFIS models for $K_x$ estimation in natural rivers. In addition, the performance of the ANFIS model was found better than that of the ANN model. A multivariate adaptive regression splines (MARS) was developed for $K_x$ estimation in rivers by Haghiabi (2016). The performance of the MARS model was compared against the ANN model and some EFs, and the modeling outcome showed that the MARS model performed better than the ANN model and the EFs in terms of accuracy. In another study, several training algorithms, including genetic algorithm (GA), imperialist competitive algorithm (ICA), bee algorithm (BA), cuckoo search (CS), and Levenberg–Marquardt (LM), were used for ANN model learning process development for $K_x$ estimation in rivers (M. Alizadeh et al., 2017). The evaluation of the predictive models showed that they could be successfully improved based on integration the swarm evolutionary algorithms. However, the performance of the CS, ICA, and BA algorithms was better than the GA and LM for training the ANN model. The use of the Bayesian network (BN) for $K_x$ estimation in natural rivers was presented by M. J. Alizadeh, Shahheydari, et al. (2017). First, the study applied the clustering technique as a data pre-processing technique to cluster the data in separate groups of similar characteristics. The study showed that the developed model was a suitable tool of pollutant transport prediction in natural rivers.

The use of evolutionary polynomial regression (EPR) model for accurate estimation of $K_x$ in rivers was presented by Balf et al. (2018). The modeling was based on the flow depth, channel width, and average and shear velocities parameters. The estimation of $K_x$ using the EPR model was compared with those estimated using other conventional estimation formulas. Based on the obtained estimation results, the introduced EPR model for $K_x$ estimation was found suitable to be incorporated in one-dimensional water quality models for better solute concentration estimation in natural rivers. An ANFIS-based principal component analysis (PCA) was performed for $K_x$ estimation by Parsaie et al. (2018). The evaluation of the model showed better accuracy of the ANFIS model compared to the experimental formulas. A study was conducted based on the Pareto-Optimal-Multigene Genetic Programming (POMGGP) equation for $K_x$ estimation (Riahi-Madvar et al., 2019). The study analyzed 503 data sets of channel geometry and flow conditions in natural streams in order to develop a hybrid model. The developed hybrid model combines the Subset Selection of Maximum Dissimilarity Method (SSMD) with Multigene Genetic Programming (MGP) and Pareto-front optimization. The combined models aimed to establish a set of selected dimensionless equations of $K_x$ and the best equation with the widest applicability. As per the authors’ findings, the proposed equation provided an accurate estimation of $K_x$ compared to the other published equations.

A hybrid predictive model based on the optimized ANN model with granular computing (GRC-ANN) was established for $K_x$ estimation in natural rivers (Ghiasi et al., 2019). The performance of the GRC-ANN model was compared with those of ANFIS/ANN models and other EFs in terms of accuracy and performance in different $K_x$ values. The results confirmed the potential of the GRC-ANN model. An integrative AI model was developed, including group method of data handling and extreme learning machine (GMDH-ELM) for $K_x$ estimation in water pipelines (Saberi-Movahed et al., 2020). The analysis of the GMDH-ELM model achieved a good level of precision in the training and testing phases. The results further showed that the proposed GMDH-ELM performed better than other AI and conventional predictive models.

Recently, several AI models, including support vector regression (SVR), Gaussian process regression (GPR), random forest (RF), M5 model tree (M5P), and multiple linear regression (MLR) for $K_x$ estimation in natural streams (Kargar et al., 2020). The study found that the M5P model with simple formulations exhibited better performance than the other AI models and EFs, and was recommended as a suitable tool for $K_x$ estimation in rivers. $K_x$ was modeled using an integrated AI model based on the Subset Selection of Maximum Dissimilarity (SSMD) method and the Whale Optimization Algorithm (WOA) as a simple optimization approach (Memarzadeh et al., 2020). The study presented a high accuracy formula for $K_x$ estimation, which was proven to be superior in terms of $K_x$ estimation compared to the existing $K_x$ estimation formulas. The WOA was also found to improve the predictive performance of the equations in other related fields by establishing the optimum coefficient values. A hybrid model was developed using hybridized ANFIS with firefly algorithm (FFA) for $K_x$ estimation (Riahi-Madvar et al., 2020). The FFA was used for the derivation of the optimum parameters of the ANFIS model. The analysis of the proposed ANFIS-FFA
model showed significant improvement compared to the standalone ANFIS model, suggesting that the parameters optimization by the nature-inspired optimization algorithms contributed significantly toward improving the generality of the ANFIS estimations. On the same manner, other scholars developed a new hybrid AI model based on the integration of Harris Hawks Optimization (HHO) algorithm with ANFIS and least square support vector machine (LSSVM) models for the $K_x$ estimation (Azar et al., 2021). The advancement of AI models pointed that there is a motivation to explore the feasibility of the emerging technology deep learning models for the hydraulic engineering simulation. Hence, a convolutional deep learning model was constructed for the $K_x$ estimation (Ghiasi et al., 2021).

**Research gap and current study motivation**

According to the reported literature review studies, researchers devoted remarkable investigations on AI models’ feasibility to improve their designs, comprehending, and evaluations for $K_x$ modeling in natural rivers. However, exploring a new version of AI models and their advanced versions is still ongoing research motivation to better understand the underlying pollutants transport in natural streams. Although several AI models have been explored for modeling the $K_x$, several issues related to the methodological aspects observed remained unsolved, such as: (i) AI models are only applicable when there are enough training samples, and this is hard to obtain in real-world scenarios; (ii) the incorporation of FS approaches characterized with non-linearity could substantially improve the feasibility of the ML models; (iii) the learning of the intricate structure of the input parameters of AI models is critical; hence, a deeper network comes with a significant increase in computational cost and overfitting owing to the associated increase in the number of parameters to train, and sometimes due to inappropriate training samples; (iv) there is still the issue of appropriate balancing of the network depth and computational efficiency, while accommodation of nonlinear FS remains an essential step.

Also, FS is known as parameter selection or attribute selection, is an independent process generally combined with the learning procedure of ML models, which aims to reduce the number of attributes to improve the performance of a predictor (Creaco et al., 2016). The central premise of FS is to remove the irrelevant or redundant features from the dataset; thus, the quality of the final estimations does not deteriorate (Meng & Li, 2017). Depending on their interaction with the ML model, FS approaches can be classified into filter models, wrapper models, and embedded models. Filter models work independently of the predictor using information from the dataset relying on a metric constructed upon several statistical tests, which can be a correlation coefficient, a distance function, or an importance measure. Wrapper models use an estimation model and select a subset of features based on the predictor performance. Although wrapper methods generally produce better performance, they are computationally intensive and time-consuming than the filter methods (Tan et al., 2020). In the embed methods, the selection of relevant features is integrated into the learning of the estimation model (Cai et al., 2018). The selected features are those that best contribute to the accuracy of the estimation model (Mirzaei et al., 2017).

**Research objectives**

This study was established to propose a ML model that integrates FS and building robust and reliable predictive model to estimate longitudinal dispersion coefficient. Despite the massive applications involving statistical and ML models reported over the past two decades, the current research proposes for the first time a Gaussian Process Regression hybridized with an evolutionary algorithm to estimate the dispersion. The motivation of the study is to provide an accurate model that automatically selects the most relevant input parameters while releasing reliable and accurate estimations. For the validation purpose, the SVR model was developed in which it has been produced good estimation in several related studies. The evolutionary search is performed by the Covariance Matrix Adaptation Evolution Strategy (CMAES), a simple, well-established, and competitive evolutionary algorithm (Hansen, 2006; Li et al., 2020).

In order to show the scope of the model, the FS step was removed from the model, and its performance on the dispersion estimation was compared with the standalone GPR and SVR models. The purpose here is to investigate that even without the selection of characteristics, the proposed model can achieve competitive results against estimation models described in the literature (Toprak et al., 2004). Then, the FS approach was included again in the model, and then extensive computational experiments showed that the proposed approach was capable of selecting the most relevant features while training the ML models, improving the performance, and producing accurate and reliable outcomes.

**Materials and methods**

**Longitudinal dispersion coefficient dataset**

The current research was established using the natural streams dataset, were collected from the published
research (Tayfur & Singh, 2005). The dataset consists of 71 geometric and hydraulic parameters observations, and longitudinal dispersion coefficients were obtained from 29 rivers in the United States.

The samples in the dataset are scaled according to:

\[ v_i = \frac{v_i}{v_{\text{max}}}, \quad v_i \in \{B, H, Q, U, u^*, U/u^*, \beta, \sigma, K_x\} \]

where \( v_i \) is the standardized value, and \( v_{\text{max}} \) is the maximum values for the samples of the considered parameter.

From the scaled dataset, 51 samples (72%) were used to compose the training modeling phase, while the remaining 20 observations samples (28%) were used for the testing phase. Tables 1 and 2 present the description of the parameters and the statistical information for training and test set, respectively. The relative shear velocity can be associated with the roughness and hydrodynamic characteristics of the river bed (Etemad-Shahidi & Taghipour, 2012). The channel shape parameter, given by \( \beta = \ln (B/H) \), reflects the vertical irregularities of the river bed (Deng et al., 2001). The channel sinuosity is defined as the channel length ratio to the valley length (Sahay, 2013).

### Feature selection encoding

The focus of this work is to develop an ML model to estimate the dispersion coefficient of natural streams. The model comprises a set of chained steps that involve processing the data, selecting the features, building the ML predictor, and finally releasing the final model to estimate the outputs. In addition, the proposed ML model aims to reduce human efforts in the simulation process when building a reliable and accurate methodological technique (Liu et al., 2017). Figure 1 depicts the hybrid predictive ML model proposed in this research integrating data processing, feature selection, and model building.

A simple procedure is used in the FS step. A feature vector, composed by 8 entries:

\[ \theta^\text{FS} = [x_B, x_H, x_Q, x_U, x_{u^*}, x_{U/u^*}, x_\beta, x_\sigma] \]

that represent the logical variables for each parameter shown in Tables 1 and 2. An entry equals 1 represent a selected feature and 0 a removed feature. For example, the feature vector \( \theta^\text{FS} = [10100001] \) means that the set \{ \( B, Q, \beta, \sigma \) \} is used in the estimation model to produce the \( K_x \) estimations. Two ML models were used along with the integrated model: Gaussian Regression Regression (GPR) and Support Vector Regression (SVR).

The parameterization of the hybrid model comprises the adjustment of two vectors \( \theta^\text{FS} \) and \( \theta^\text{MB} \). These two vectors are then concatenated in a parametric vector

\[ \theta = [\theta^\text{FS}, \theta^\text{MB}] \]

where \( \theta^\text{FS} \) is the feature vector, and \( \theta^\text{MB} \) is the vector of the internal parameters of the ML models.

### Gaussian processes regression

Given a dataset of observed samples \( \{x_i, y_i, i = 1, 2, \ldots\} \), a Gaussian Process (GP) is a non-parametric approach in which to find a distribution over the possible functions \( f(x) \) that are consistent with the samples. In Gaussian process regression (GPR), it is assumed at any point \( x \) is assigned the output \( \hat{y} \) of a random variable \( f(x) \) that can be written as:

\[ \hat{y} = f(x) + \varepsilon \]

where \( \varepsilon \) is an independent and identically distributed noise contribution. The predictive distribution of the output parameter \( \hat{y} \), given its input \( x \), is given by (Wang & Chen, 2015):

\[ \hat{y}(x) = k^T(x)(K + \alpha I)^{-1}y, \]

where \( y = [y_1, y_2, \ldots, y_N]^T \) is the vector of training data, \( K = [K(x, x_1), K(x, x_2), \ldots, K(x, x_N)]^T \) is a positive definite kernel function, \( K \) is the covariance matrix function, and \( \alpha \) is a regularization parameter.

The covariance function models the dependence between the function values at different input points \( x_i \).
and $x_j$ (Schulz et al., 2018):

$$K_{ij} = K(x_i, x_j) = k_0 \frac{2^{1-v}}{\Gamma(v)} \left( \frac{2 \sqrt{\nu}}{l} \right)^v \times K_v \left( \frac{||x_i - x_j||}{\sqrt{2 \nu l}} \right)$$

where $v$ and $l$ are positive parameters, $\Gamma(\cdot)$ is the Gamma function and $K_v(\cdot)$ is the modified Bessel function (Kumar, 2020). The building of the GPR model involves the proper setting of the kernel parameters $k_0, v, l$ and the regularization coefficient $\alpha$. The model building vector for GPR ML model is represented by $\theta^\text{MB} = [k_0, v, l, \alpha]$.

**Support vector regression – SVR**

The Support Vector Regression (SVR) builds a linear regression model of the form:

$$f(x) = \sum_{j=1}^{N} w_j K(x, x_j) + b$$

where $K(\cdot, \cdot)$ is a nonlinear transformation, $w = [w_1, \ldots, w_N]$, is the vector of weights, $b$ is a constant threshold and $N$ is the number of samples. The nonlinear transformation used in this paper is the radial basis kernel function of the form

$$K(x, x_i) = \sum_{i=1}^{N} \exp(-\gamma ||x - x_i||^2)$$

where $\gamma$ is the bandwidth parameter.
In SVR the optimal $w$ and $b$ are computed by minimizing the functional (Kargar et al., 2020):

$$ J = \sum_{i=1}^{N} w_i^2 + \frac{C}{N} \sum_{i=1}^{N} L_\varepsilon (y_i - f(x_i)),$$

$$ L_\varepsilon (y - f(x)) = \begin{cases} 0 & \text{if } |y - f(x)| \leq \varepsilon \\ |y - f(x)| & \text{otherwise} \end{cases}$$

where $y_i$ is output data associated with $x_i$, $L_\varepsilon$ is the $\varepsilon$-insensitive loss function (Gunn, 1998) and $\varepsilon$ is a SVR parameter. The internal parameters of the SVR model to be adjusted are $C, \varepsilon$ and $\gamma$, resulting the the model building parameter vector $\theta_{MB} = [C, \varepsilon, \gamma]$.

The kernel function in Equation (4) assists in constructing a nonlinear decision hypersurface on the SVR input space, while the bandwidth $\gamma$ acts as the inverse of the variance of the kernel function. The regularization parameter $C$ determines the relationship between the cost of minimizing the training error and minimizing the model’s complexity. The tube size of $\varepsilon$-insensitive loss function controls the accuracy placed on the training samples (Chen, 2007).

**Covariance matrix adaptation evolution strategies**

As shown proposed in the framework shown in Figure 1, an evolution strategy assists the FS and the adjustment of the internal parameters of the ML model. The encoding for a solution implementing the Gaussian Process is represented by the 12-dimensional vector

$$ \theta = [x_B, x_H, x_Q, x_U, x_{Q^*}, x_{U^*}, x_B, x_H, x_Q, x_U, x_{Q^*}, x_{U^*}, k_0, \nu, l, \alpha]$$

while the 11-dimensional parametric vector

$$ \theta = [x_B, x_H, x_Q, x_U, x_{Q^*}, x_{U^*}, x_B, x_H, x_Q, x_U, x_{Q^*}, C, \varepsilon, \gamma]$$

encodes a solution implementing a Support Vector Regression model.

The hybrid model implements the Covariance Matrix Adaptation Evolution Strategy (CMAES), a variant of the evolution strategy algorithm (Hansen & Ostermeier, 2001). CMAES performs the optimization by iteratively sampling a population of $\lambda$ candidate solutions from a multidimensional normal distribution based on a vector of means $\bar{m}$, a covariance matrix $C$ using a step size $\sigma$ (Beyer & Schwefel, 2002). Figure 2 shows the pseudocode of the CMAES algorithm.

CMAES starts the search creating a population of $\lambda$ candidate solutions. Each solution is then evaluated using the objective function given by the Root Mean Discrepancy averaged over all $N$ observations

$$ \text{RMD} = \frac{1}{N} \sum_{i=1}^{N} \text{DR}_i^2 = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{K_{xp_i}}{K_{xmi}} \right)^2 $$

where

$$ \text{DR}_i = \log(K_{xp_i}/K_{xmi}) $$

is the Discrepancy Ratio (DR) for the $i$th sample, $K_{xp_i}$ and $K_{xmi}$ are the estimated and observed dispersions. The Discrepancy Ratio (DR) is used as error measure for each sample. When $K_{xp} = K_{xmi}$, DR is equaled to zero, and there is an exact estimation. Otherwise, there is either an over-estimation when DR > 0, that means $K_{xp} > K_{xmi}$, or an under-estimation which leads to DR < 0 and consequently $K_{xp} < K_{xmi}$.

The 5-fold cross-validation approach is employed to assess the fitness of the individual. This technique randomly splits the dataset into five subsets. The ML model using the parameters $\theta_{MBi}$ is trained in four of them, and the RMD from Equation (5) is calculated in the remaining subset. This procedure is repeated five times, and the objective function associated with the candidate solution is the RMD averaged over the five folds.

Then, CMAES evolves the population over $N_G$ generations. The goal is to find the best parameter $\theta^*$ such that the RMD from Equation (5) to be reduced to the lowest possible value. At each iteration $k$, the algorithmic parameters $m_k$, $C_k$ and $\sigma_k$ are modified based on the current evaluated population according to the algorithm shown in Figure 2. A detailed explanation of the CMAES algorithm can be found in Li et al. (2018).

**Models assessment**

After the simultaneous optimization of the feature and model building vectors, the ML model is ready to be used as a $K_x$ predictor in the test set. Three metrics were used to assess the model performance: Root Mean Square Error (RMSE), Coefficient of Determination ($R^2$) and Accuracy (Yaseen, 2021).

The RMSE is calculated according to

$$ \text{RMSE} = \frac{1}{N} \sum_{i=1}^{N} (K_{xmi} - K_{xp_i})^2 $$

where $N$ is the number of samples, $K_{xmi}$ and $K_{xp_i}$ are the measured and estimated longitudinal dispersions for the $i$th sample, respectively. The Coefficient of Determination is expressed as

$$ R^2 = \frac{\sum_{i=1}^{N} (K_{xp_i} - \bar{K}_{xmi})^2}{\sum_{i=1}^{N} (K_{xmi} - \bar{K}_{xmi})^2} $$
where $\bar{K}_{xm}$ is the mean of the observed longitudinal dispersions.

The Accuracy is categorized by the number of DR values between $-0.3$ and $0.3$ relative to the total number of samples (Tayfur & Singh, 2005):

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} I(\text{DR}_i \leq 0.3)$$ (9)

where $\text{DR}_i = \log(K_{xp}/K_{xm})$ is the Discrepancy Ratio for the $i$th sample, given by Equation (6), $K_{xp}$ is the estimated longitudinal dispersion, $K_{xm}$ is the measured longitudinal dispersion, and $I(\cdot)$ is a indicator function that returns 1 if the condition is true and 0 otherwise.

**Computational procedure description**

A robust procedure was proposed to report the results: (i) perform 100 independent runs establishing a unique random seed for each run; (ii) for each run, the hybrid models built on the training set are used to estimate the dispersion coefficients in the test set; (iii) calculate the RMSE metric and Accuracy to assess the performance of each model. The selected features in the first step of the experiments are summarized in Table 3.

### Table 3. The input features associated with established models.

| Case   | $\theta^{FS}$ (fixed mask) | Active features |
|--------|----------------------------|-----------------|
| 0      | 111111111                   | $U, H, B, u^*, Q, U/u^*, \beta, \sigma$ |
| 1      | 111000000                   | $U, H, B$       |
| 2      | 000010000                   | $Q$             |
| 3      | 100000000                   | $U$             |
| 4      | 100000010                   | $U, \beta$     |
| 5      | 100000111                   | $U, \beta, \sigma$ |
| 6      | 100001000                   | $U/u^*$         |
| 7      | 000011111                   | $U/u^*, \beta, \sigma$ |

The computational experiments were divided into three parts:

1. **Modeling the $K_x$ without feature selection.** Performing the experiments using the proposed ML model by considering all features (Case 0) and also reproduce the models presented in (Tayfur & Singh, 2005) (Cases 1–7) shown in Table 3. The vector $\theta^{FS}$ is kept constant for these experiments, depending on the $K_x$ model. Only the ML parametric vector $\theta^{MB}$ is adjusted to fine-tune the ML model in this experiment.

2. **Modeling the $K_x$ using evolutionary FS.** The proposed hybrid model is shown in Figure 1 to simultaneously search for the most suitable feature set and...
the model parameters. The FS procedure allows for exploring combinations of parameters that are not presented in Table 3, and that can potentially produce better predictability performance when those inputs parameters supplied for the GPR and SVR models.

(3) Modeling the $K_x$ with the most frequent established features. Building the ML models using the most frequent features found in the previous experiment exploits previous knowledge on the features and model parameters to propose an accurate model.

Considering the input combinations presented in Table 3, Case 0 employs the original features to model the dispersion coefficient of natural streams. Cases 1–7 were proposed in (Tayfur & Singh, 2005), where a neural network was used to estimate $K_x$. Case 1 considers three parameters involving flow and geometric characteristics: the flow velocity $U$, the flow depth $H$, and the channel width $B$. Case 2 employs only flow discharge $Q$ as the input parameter. As the velocity plays an essential role in the pollutant spill, Case 3 considers only flow velocity $U$. Case 4 considers flow velocity $U$ and channel shape parameter $\beta$ as input parameters, while Case 5 considers channel sinuosity $\sigma$ in the feature vector $\theta_{FS}$ along with the channel shape parameter $\beta$ and flow velocity $U$. Case 6 considers only the relative shear velocity $U/\mu^\ast$. Finally, Case 7 considers the channel shaper parameter $\beta$ and channel sinuosity $\sigma$ along with the relative shear velocity $U/\mu^\ast$ as input features.

**Uncertainty analysis**

Uncertainties in the dispersion values can occur due to inadequate estimation procedures and loss of tracker or measurements made in the advective zone (Moreno-Rodenas et al., 2019; Sattar & Gharabaghi, 2015). Other factors, such as software and hardware errors, can also interfere with the measurement of the hydraulic characteristics of a river (Rutherford, 1994). Consequently, the dispersion coefficients and hydraulic characteristics may present some uncertainties due to the reasons described above. This subsection presents a quantitative assessment of the uncertainties of the developed models in predicting longitudinal dispersion. The uncertainty analysis is applied to the testing phase dataset, comparing the predictive models’ capacity.

The uncertainty analysis is calculated using the 95% confidence band around the estimated value $K_{xmj}$. The confidence band can be approximated in the range

$$(K_{xmj} \times 10^{-\bar{\varepsilon}+1.96S_e}, \ K_{xmj} \times 10^{-\bar{\varepsilon}+1.96S_e})$$

where $K_{xmj}$ is the measured dispersion, $K_{xpj}$ is the estimated dispersion for the $i$th sample, and $\bar{\varepsilon}$ and $S_e$ are the mean and standard deviation of the logarithm errors, respectively. The mean and standard deviation of the estimation are calculated according to

$$\bar{\varepsilon} = \frac{1}{N} \sum_{j=1}^{N} \varepsilon_j \quad \text{and} \quad S_e = \sqrt{\frac{1}{N-1} \sum_{j=1}^{N} (\varepsilon_j - \bar{\varepsilon})^2} \quad \text{where}$$

$$\varepsilon_j = \log_{10}(K_{xpj}) - \log_{10}(K_{xmj}) \quad j = 1, \ldots, N$$

and $N$ is the number of test samples.

**Monte Carlo simulation of the estimated dispersion values**

To assess the effect of the stochastic feature of stream geometric and hydraulic parameters and their impact on estimated dispersion values, a Monte Carlo Simulation (MCS) method was implemented for finding the approximate distribution of the $K_x$ outcomes of the hybrid models. MCS is a robust and easy-to-implement numerical technique to calculate the output uncertainty of a model and assessing the distributions of the outputs that would result from the joint distribution of input uncertainties (Chaudhary & Hantush, 2017; Vose, 1996).

In the experiments conducted in this research, the variations of the input parameters were modeled using the uniform distribution with the lower and upper bounds given by the minimum and maximum values for each parameter, as shown in Tables 1 and 2. For each MCS run, the deterministic outcome is calculated for each model developed in this study. A total of 250,000 outcomes were calculated for $K_x$. The Mean Absolute Deviation (MAD) around the median of the output distribution is written as:

$$\text{MAD} = \frac{1}{250000} \sum_{i=1}^{250000} |K_{xp} - \text{median}(K_{xp})|$$

while the uncertainty of the model output can be given by:

$$\text{Uncertainty} \% = \frac{100 \times \text{MAD}}{\text{median}(K_{xp})}$$

where $K_{xp}$ is the estimated dispersion for the $i$th sample. A detailed description of the MCS method can be found in (Sattar & Gharabaghi, 2015).
**Computational experiments and discussion**

This section covers all the attained modeling results of the proposed ML model to estimate the longitudinal dispersion coefficient. All the computational process, FS scenarios, and the validation against the literature review (Tayfur & Singh, 2005) (i.e. ANN model ‘reference model’) were reported in this section.

The experiments were conducted on a cluster of computers with the following specifications: Intel Xeon E5620 (2 cores of 2.4 GHz, and 12MB cache memory), RAM of 16 GB, and operating system Linux Linux 2.6.32.

During the optimization steps of CMAES, the algorithm samples a general multivariate normally distributed random vector that has a complexity of $O(d^2)$, where $d$ is the number of dimensions of the candidate solutions. Updating the covariance matrix has a complexity of $O((\mu + 1)d^2)$, where $\mu$ is a CMAES parameter (Ros & Hansen, 2008). As a result, the steps of CMAES algorithm have a computational complexity of $O(d^2)$ where $d$ is the number of dimensions of the candidate solutions. Concerning the ML models, the training procedure for GPR requires solving a linear system, while SVR training involves solving a quadratic programming problem (Chen, 2007). GPR and SVR algorithms have $O(N^3)$ time complexity and $O(N^2)$ memory complexity, where $N$ is a size of the training set (Belyaev et al., 2014).

**Modeling $K_x$ without feature selection**

Figure 3 compares the results of the $K_x$ modeling without FS approach. In this figure, the results of GPR and SVR models were averaged over 100 independent runs. According to Figure 3, it can be observed for Case 0, the average value of RMSE is considerably lower than those produced by the reference model for both GPR and SVR models. The averaged accuracy obtained with the GPR model was competitive with those produced by the reference model. However, there was a decrease in the performance for the accuracy of the SVR model. Notably, for the case where the width-to-depth ratios are lower than 50, the RMSE ($B/H < 50$) values were remarkably better than the reference model, representing a reduction of 64.9% for the GPR and 67.8% for SVR. This finding shows that the standalone ML models can provide accurate estimations for narrow streams. On the other hand, they did not produce better results than the reference model when the extreme values were not considered, as it can be observed for the RMSEs less than 100 m$^2$/s, RMSE ($K_x < 100$).

The modeling results for Case 1 were interesting since the features used as input parameters in this model are a subset of the collected ones. The parameters used in this model, $U$, $H$ and $B$, are those exhibited higher correlations with the dispersion coefficient as it can be seen in Figure 3.
Figure 4. Correlation coefficients using Spearman rank correlations between the eight input parameters and the longitudinal dispersion coefficient.

Figure 4. Figure 5 shows the strength of the linear relationships for $U$, $H$ and $B$. It is important to notice that, although informative, the linear correlation may not represent the complex and the nonlinear nature of $K_x$ (Noori et al., 2016).

Considering the reported results in Figure 3, although the GPR model has produced better accuracy, the SVR model produced smaller RMSE when compared to the GPR model and the reference model. The GPR model poorly estimates the dispersion coefficient for narrow channels, as observed in the higher values for the RMSE if $(B/H < 50)$ compared to the SVR model. However, considering the RMSE $(B/H < 50)$, both models produced better estimations than the reference model.

For Cases 2, 3, and 6, the standalone models performed inadequately. It is interesting to notice that Case 2 involves the parameters implicitly in Case 1 since the flow discharge is the product of flow depth, velocity, and channel width, $Q = HU/B$. In addition, the linear correlation between the flow discharge and the dispersion can be discarded due to the small correlation as shown in Figure 4. The models for Cases 2, 3, and 6 were built upon one parameter, which is insufficient to represent the nonlinear relationship between the single feature and the dispersion coefficient. The finding corroborates the results described in Tayfur and Singh (2005), that solely the flow discharge $(Q)$, flow velocity $(U)$, and the relative shear velocity $(u^*)$ are not sufficient to estimate the nonlinear behavior of the dispersion coefficient. Case 4 was built upon Case 3, including the channel shape parameter $(\beta)$ for Case 4. Whereas in Case 5, the channel sinuosity was added to Case 4. The standalone (i.e. GPR and SVR) models performed similarly to the conducted ANN model ‘reference model’ developed by Tayfur and Singh (2005). However, it was not able to improve the results when compared to the previous cases. Case 7 considered the relative shear velocity $(U/u^*)$ as in Case 6, and channel shape parameter $(\beta)$ along with the channel sinuosity $(\sigma)$ to estimate the dispersion coefficient. The statistical metrics shown in Figure 3 demonstrated the inclusion of the channel shape parameter, and the channel sinuosity did not improve the $K_x$ estimations.

The results depicted in Figure 3 obtained after several independent runs allow for draw the following conclusions on the test Cases presented in Table 3. From Cases 0-7, GPR and SVR models have produced the overall best results for Case 0 and Case 1, as it can be seen in Figure 3. Given that the GPR and SVR models’ performance observed the highest for Cases 0 and 1, the comparative analysis will be focused at forward on these two cases. The RMSE for both models was inferior to the ANN model developed in Tayfur and Singh (2005). In Case 0, where the eight parameters were considered, the GPR and SVR models showed an average reduction of 58.9% and 58.8% using the RMSE metric, respectively. Similar behavior was observed for the narrow channels where $B/H < 50$: the reduction was 35.1% and 32.2% for GPR and SVR models, respectively. In addition, the average Accuracy for GPR was only 3.24% small than the reference model.

In Case 1, the averaged RMSE values obtained using GPR and SVR models were smaller than the RMSE attained by the reference model. However, the RMSE produced by GPR model was higher than that provided by SVR model. Similar behavior is observed for the RMSE of narrow channels $(B/H < 50)$. For the Case 1, the GPR
model produced an average Accuracy of 70.09% against 75% for the reference model, resulting in a percentage difference of 6.55% on average.

For Cases 0 and 1, Figure 6 depicts the scatterplots of the measured and estimated dispersion coefficient for the best models over the test set (according to RMSE). Besides the RMSE, accuracy and the coefficient of determination associated with each model were displayed as well. Although a few estimated $K_x$ present a visible error, the estimations for GPR and SVR models revealed a good agreement with the observations, as it can be verified by the $R^2$ magnitudes, above 0.90 for GPR in Cases 0 and 1. For SVR, $R^2 = 0.93$ for Case 0 and $R^2 = 0.89$, were attained for Case 1.

In the computational experiments, a total of 100 independent runs were performed. As a consequence, it is interesting to analyze the distribution of the internal parameters in all executions. As Cases 0 and 1 produced better results than the other ones, the discussion is focused on these cases. Figure 7 shows the distribution of the internal parameters for Cases 0 to 7. The GPR model parameters are namely $k_0$, $\nu$, $l$, and $\alpha$. The first three parameters control the shape of the kernel shown in Equation (3) used to compute the estimations, while $\alpha$, which is added to the diagonal of the kernel matrix during fitting, rules the level of noise in the observations. The $k_0$ parameter controls the magnitude of the GPR model approximation, and the final solutions showed similar distributions in cases 0 and 1. Similar behavior was noticed for the parameter $\alpha$, which is associated with noise in the observations. The dispersion is a measure associated with a natural process, and thus that noise becomes an essential factor in the estimates. The parameter $\nu$ is a critical parameter in the kernel function and it controls the GPR smoothness (Stein, 1999): the smaller $\nu$ value, the less smooth approximated solution achieved. The distribution of $\nu$ in Figure 7 shows the GPR estimation functions produced for Case 1 smoother than Case 0. In addition, the length scales $l$ for Case 1 are higher than Case 0 as well. These combined circumstances potentially

![Figure 6. Scatter plots for the best models according to Root Mean Squared Error (RMSE).](image-url)
led to the highest RMSE values for case 1, as it can be visualized in Figure 3.

Figure 8 displays the internal parameters distribution of the SVR model including $C$, $\gamma$, and $\varepsilon$. $C$ is the penalization parameter, $\varepsilon$ specifies the penalization associated with the training loss function, and $\gamma$ is kernel coefficient in Equation (4). By analyzing the boxplots, it can be observed that the distribution of the parameters $\gamma$ and $\varepsilon$ were very similar for both cases. On the other hand, the values of $C$ parameter for Case 1 were higher than those found for Case 0. In the support vector machine model, $C$ plays an important role as a regularization parameter. The regularization strength is inversely proportional to $C$. As a result, the SVR estimations for Case 0 were smoother than those produced for Case 1. The smaller RMSE values for Case 1 shown in the barplots in Figure 8 (averaged in 100 runs) support this interpretation.

When compared to the reference model, it can be noticed that GPR and SVR models presented a lower performance in predicting the dispersion coefficient for the extreme values of $(K_x > 100)$, which are discarded for Cases 0 and 1. One possible explanation for the depreciated performance of RMSE for small $K_x$ is that those models do not suitably explore the relationship among the features for small dispersion values. For case 0, the interaction among the eight features might not be beneficial to estimate small $K_x$ values. In contrast, for Case 1, the features $U$, $B$, and $H$ that conserve linear relationship with the dispersion coefficient might not be enough to represent the non-linearities of the dispersion coefficient. Considering that the evolutionary strategy appropriately
determines the parameters $\theta_{\text{MB}}$ of the models, an alternative is to choose a set of input parameters that balance the predictive resources with extreme values. Likewise, this set of parameters should produce estimates with low RMSE values at high precision.

**Modeling $K_x$ with evolutionary feature selection**

Considering the experiments in the previous subsection, it is not clear whether other combinations of parameters can lead to better predictability performance. Despite the explanation provided by the authors in their study (Tayfur & Singh, 2005), it can be argued that the parameters associated with the cases shown in Table 3 were chosen, to their major extent, based on the correlation with the longitudinal dispersion coefficient. Figures 4 and 5 show that the parameters with the highest correlations are those that comprise the sets in Table 3. For example, the $U$ parameter appears in 5 out of 8 cases. Besides, the parameters $U$, $B$, and $H$ are part of Cases 1 and 2, which produced the best-averaged results.

The hybrid model developed in the current research can search for the most suitable predictor parameters to explore combinations that were not previously considered. It can be noticed that due to the nonlinear natures of the predictors GPR and SVR models, a careful choice of parameters can produce an improved performance of the ML models. Also, parameters with slight correlations with $K_x$ might positively impact the final estimations. The subsequent experiments aim to analyze the impact of the evolutionary FS on assessing the $K_x$ estimation performance metrics.

Table 4 summarizes the results obtained for GPR and SVR hybrid models with FS shown in Figure 1. The first column shows the cases, and the second column displays the ML model. The RMSE is shown in the third column, while the Accuracy appears in the fourth column. The fifth two columns show the RMSE when extreme values are discarded, and the last column displays the RMSE for narrow streams. Analyzing the reported results in Table 4, it is possible to observe that the FS step on the model has contributed considerably to improve the ML models’ metrics. In addition, Table 4 presents the results for Cases 1 and 2 to compare the performance of the models when the FS was not applied. Overall, the RMSE values decreased for both GPR and SVR models. As observed for the models without implementing FS, GPR-FS and SVR-FS models produced smaller RMSE than the reference model. On average, the GPR-FS model achieved an improvement of 16.5% when compared to Case 0 and 50.2% concerning Case 1. Whereas the improvement achieved by the SVR-FS model to Cases 0 and 1 were 21.0% and 21.6%, respectively.

Regarding the Accuracy, the SVR-FS model achieved a slight improvement over the SVR model. On contrast, the GPR-FS model achieved a consistent improvement compared to the standalone model without FS. In particular, the averaged Accuracy is the highest among all models. When the extreme dispersion coefficients ($K_x > 100$) values were excluded, there was no improvement in the RMSE for both models. However, the evolutionary hybrid model implementing GPR and SVR models produced an acceptable predictability performance as per the evaluation metrics for the narrow streams. In particular, the average values of RMSE achieved by the SVR-FS model can be highlighted, as it can be seen in the last column of Table 4. Figure 9 depicts the scatter plots for the best GPR-FS and SVR-FS models, that were chosen according to the smaller RMSE. Considering GPR-FS model, it can be observed that RMSE decreased while Accuracy increased when comparing the scatter plots from Figures 6 and 9.

The analysis of the forgoing results supports the conclusion that incorporating the FS procedure in the ML development learning process leads to a considerable improvement in predicting the dispersion coefficient. The proper selection of input parameters is beneficial for the model as this interaction increases the accuracy while decreases the averaged errors.

The results in Table 4 can help assess the models’ performance improvement. However, they do not show which parameters led to these gains. Figure 10 shows the distribution of the selected features after the training process of the evolutionary hybrid model. From this figure, one can observe that GPF-FS and SVR-FS models chose a total of 13 sets of features at the end of the evolutionary search procedure. It was observed that the parameters selection process is heuristic, and there is no guarantee of choosing the same set of parameters in all runs. The barplot on the left side shows the results of GPR-FS model, the set of $B$, $U$, $\alpha$, $Q$ parameters has been selected 24 times, the $B$, $U$, $\sigma$ was selected 18 times and $B$, $U$, $\alpha$, $\beta$ 13 times. These three sets were selected 55 out of 100 times, and they share the parameters $B$, $U$,

### Table 4. Hybrid evolutionary approach with FS.

| Cases | Model       | RMSE  | Accuracy  | RMSE ($K_x < 100$) | RMSE ($B/H < 50$) |
|-------|-------------|-------|-----------|--------------------|-------------------|
| Case 0 | GPR         | 99.70 (29.05) | $67.91 (5.05)$ | $31.80 (3.97)$ | $63.13 (10.63)$ |
|       | SVR         | 112.34 (26.40) | $57.73 (7.06)$ | $41.23 (4.28)$ | $58.80 (7.58)$ |
|       | REF         | 193.00 (–)    | $70.00 (–)$    | $19.30 (–)$    | $183.00 (–)$    |
| Case 1 | GPR         | 167.34 (15.45) | $71.05 (6.19)$ | $41.28 (8.16)$ | $114.68 (98.36)$|
|       | SVR         | 113.22 (35.35) | $62.59 (5.96)$ | $45.18 (8.84)$ | $61.73 (14.43)$ |
|       | REF         | 193.00 (–)    | $75.00 (–)$    | $21.20 (–)$    | $183.00 (–)$    |
| Hybrid | GPR-FS      | 83.26 (26.41) | $76.73 (6.60)$ | $31.25 (5.18)$ | $52.41 (14.61)$ |
|       | SVR-FS      | 88.78 (13.66) | $65.77 (8.99)$ | $39.51 (8.42)$ | $46.10 (7.21)$ |

Note: The best results are shown in boldface with standard deviation within parentheses. Values indicated with – were not available at the source.
and \( \sigma \). A similar analysis for the SVR-FS model in the bar graph of Figure 10 shows that \( B, U, \sigma, \beta \) was chosen 33 times, \( B, U, \sigma, \beta \) 20 times, and \( B, U, \sigma \) 12 times. These three sets were selected 74 times throughout 100 independent executions, approximately two-thirds of the total runs.

Figure 11 shows that \( U, \sigma, B, Q \) were the most frequent parameters for the GPR-FS model, in that order. For the SVR-FS model, the most frequent features were \( U, \sigma, B, \beta \), in this order as well. Moreover, for both models, the \( U \) and \( \sigma \) parameters were select in all runs, while \( B \) appears 98 times for the GPR-FS model and 96 times for the SVR-FS model. From most frequent parameters, \( U \) and \( B \) have the strong correlation with the dispersion coefficient, whereas river sinuosity \( \sigma \) has the lowest correlation, as it can be seen in Figure 4. The FS has shown that the sinuosity plays an essential role in the context of the hybrid model, suggesting \( \sigma \) holds a solid nonlinear relationship with \( K_x \) that was not captured by linear correlation approaches.

Some literature studies discussed that sinuosity is related to river meandering (Deng et al., 2002) other exogenous factors such as vegetation (Camporeale & Ridolfi, 2010; Savickis et al., 2016) and topography (Timar, 2003; van Dijk et al., 2013), that also affect the dispersion coefficient. As shown in the results presented in this study, the incorporation of sinuosity is beneficial for the adopted ML models performance.
Figure 11. Hybrid model with evolutionary FS approach. Number of times each parameter appears in the active sets over 100 runs.

**Modeling $K_x$ using the most frequent features found by evolutionary selection**

The computational experiments carried out in this subsection assist in determining the importance of the parameters indirectly. This importance takes into account their impact on the hybrid ML models. The results shown in Table 4 and Figure 10 suggest that some parameters constitute the core that improves the learning processes of the applied ML models. A detailed analysis of the frequency that the parameters appear in the selected sets is presented in Figure 11. Through this figure, it is possible to verify that the $\sigma$ parameter appears in all selected sets. The same occurs with the $U$ parameter.

Based on this analysis, the essential parameters can be determined as $U$, $B$, and $\sigma$, which can be interpreted as the best selection of the proposed evolutionary FS approach. These three parameters are used to propose Case 8 shown in Table 5.

Table 6 presents the prediction accuracy improvement for all metrics. In particular, it can be observed the improvement for the GPR model based on the Accuracy and RMSE for dispersion coefficients below 100 m$^2$/s. Figure 12 displays the scatter plots for the best models of Case 8 according to the RMSE. The scatter plots showed the models yielded acceptable results accuracy. This behavior is expected because careful selection for the significant parameters represents the indirect introduction of specific knowledge about the features. Besides, this specific knowledge could only be obtained after analyzing several simulations carried out with the evolutionary selection model. The most frequent parameters allow capturing the dispersion coefficient’s essential behavior, allowing the effective modeling of the $K_x$ by the ML models. However, when this knowledge is not available before carrying out the estimations, the FS step in the integrated model is indispensable to construct an accurate low-complexity model.

**Uncertainty analysis**

Table 7 summarizes the results of the uncertainty estimation. The first column shows the model’s acronyms or the reference where they were collected. The second column displays the case, according to Table 3. The third column shows the models’ mean estimation errors, while the last two columns display the width of the uncertainty band and the 95% estimation interval error. According to the results presented in this table, all models overestimated the $K_x$ values as revealed by positive values of mean estimation errors, except for SVR in Case 7.

Following the GPR model’s attained results, the width of uncertainty bands for the hybrid models vary considerably, ranging between 0.25 and 0.68. However, for Cases 0, 1, 6, 7, and 8, their width bands are smaller than those presented by the reference models. The GPR-FS model equipped with automatic FS presented comparable uncertainty bands equals 0.29. The uncertainty bands for the SVR model presented similar results: Cases 0, 1, 5, 6, 8, and SVR-FS model exhibited smaller uncertainty bands compared to the reference models. Moreover, GPR and SVR models presented competitive for the lowest 95% confidence estimation error interval. The results presented in Table 7 showed that the GPR models (Cases 0, 1, 6, 7, 8 ans FS) and SVR models (Cases 0, 1, 5, 6, 8 ans FS) produced better performance with respect to model estimation uncertainty band and 95% estimation interval.

Uncertainty analysis results for the $K_x$ using the developed models were presented in Table 8. The uncertainty analysis shows that the MAD values for both GPR and
SVR models were smaller than the reference models that appear in the last six rows of Table 8. The uncertainties for GPR and SVR models were similar to those calculated for the reference models.

Figure 13 shows the scatterplot of Uncertainty (%) and RMSE for all models. Each pair \((x, y)\) represents a particular model, where the models GPR, GPR-FS, SVR, and SVR-FS are depicted in different colors. The size of the point represents the number of features in the model. Cases 0-8 are shrunk to C0-C8, and FS means FS to facilitate the visualization. Figure 13 helps to simultaneously compare the performance on the estimation of dispersion and the uncertainty of the models. From this figure, it can be observed that the GPR (Case 2) produced small uncertainty; however, this model shows a higher RMSE. It should be noticed, this model uses one parameter \(Q\), as shown in Table 3, which may have contributed to the reduced model's accuracy. Conversely, SVR (Case 0) presents a small RMSE but with higher
Uncertainty. Cases 0, 1, 8, and FS produced RMSE below 100 m$^2$/s for GPR and SVR. The following models presented the smaller RMSE and Uncertainty: GPR-FS, GPR (Case 8), GPR (Case 1), and GPR (Case 0). Eventually, it can be concluded that it is not possible to establish a clear relationship between the number of parameters in the model with the observed RMSE or Uncertainty. Despite the limited number of models developed in this research, this figure shows a trade-off between RMSE and Uncertainty, which suggests them as conflicting objectives in the context of $K_x$ estimations.

Modeling verification against the literature formulations and ML models

In order to demonstrate the capacity of the proposed predictive models, it is essential to validate those models against established models over the literature. Table 9 presents the validation of the current research proposed models (i.e. GPR, SVR, and GPR-FS) and the existed EFs and the conducted several ML models over the literature. By considering the coefficient of determination as a comparative metric. The EFs demonstrated a low estimation accuracy. On the other hand, ML models were exhibited better productivity performance in comparison with the EFs, however, with some restricted accuracy. Based on the exhibited literature review modeling results, the attained estimation capability of the developed GPR, SVR, and GPR-FS models revealed better accuracy with a maximum $R^2 = 0.95$. It is worth mentioning here, although the predictive models were developed based on a limited number of observations (71 samples), the current research evidenced the capability of the developed models on mimicking the actual relationship between the geometric and hydraulic river system parameters and the $K_x$.

Model strengths, limitations and further research

The obtained results and the reported analysis in this paper showed that the proposed evolutionary model in this paper contributes to the development of an automated computer aid model. The model has demonstrated the ability to automatically identify the most relevant features in the dataset and thus initiated the reliable learning process for the predictive model. As a result, the model can gradually learn to focus on critical data attributes during the model building approach.

The size of the dataset imposes a limitation on the proposed approach in this paper. The dataset consists of hydromorphological parameters collected in the United States. Applying the evolutionary hybrid model to other rivers on other continents, even with similar characteristics, must be done with care. In this context, it is critical to add more samples to the dataset. However, data acquisition involves a laborious and time-consuming experimental data collection, which represents a payoff for the modeling process.

For future research, the possibility to develop alternative approaches by combining two or more ML models in an ensemble fashion is to be investigated (Elder, 2018; Kotu & Deshpande, 2019). This is for the purpose of exploiting the estimation capabilities of different ML models. Nevertheless, the resulting model can become quite complex in this context, with a potentially large number of parameters. Additional research involves implementing evolutionary multiobjective search algorithms (Jimenez et al., 2017), to reduce the number
of model parameters and search for the most helpful input parameters while maximizing the estimation performance.

**Conclusions**

The dispersion of pollutants in rivers involves complex phenomena, requiring simplifications and assumptions, leading to drawbacks to accurate estimations. In this context, ML models arise as an alternative to dispersion coefficient modeling. However, the performance of these models depends on the proper input parameters selection. In this research, a new hybrid ML model was proposed that embeds a simple approach for the associated parameters selection for modeling the dispersion coefficient of natural streams. An evolution strategy assisted the task of adjusting the selected parameters. Based on the reported literature, there is no attempt to integrate the FS approach with the estimation model developed to determine the dispersion coefficient. In general, the FS task was performed in a separate step and did not consider the estimation model. The performance of the models was measured based on the RMSE and Accuracy. The research findings are summarized as follows:

* The proposed predictive model in this research automatically searches for the essential parameters and the ML models’ internal parameters.
* The modeling results showed that with the incorporation of the FS approach, an efficient methodology was established for the dispersion coefficient simulation.
* Based on the proposed evolutionary FS approach analysis, the essential parameters for modeling the $K_x$, were identified as $U$, $B$, and $\sigma$.
* A limited number of geometric and hydraulic river system parameters (71 Observations) were effectively constructed for the $K_x$ determination.
* The uncertainty analysis indicated that GPR and SVR models presented competitive results for the lowest 95% confidence estimation error interval.
* Overall, the research confirmed the feasibility of the hybrid ML model (i.e. GPR-FS) for modeling the longitudinal dispersion coefficient in natural streams.

**Acknowledgements**

The authors appreciate the supports received by the Deanship of Scientific Research at King Khalid University under Grant number RGP 1/372/42.

**List of acronyms**

| **AI** | Artificial Intelligence |
| **ANN** | Artificial Neural Network |
| **ANFIS** | Adaptive Neuro-Fuzzy Inference System |
| **BA** | Bat Algorithm |
| **BN** | Bayesian network |
| **CMAES** | Covariance Matrix Adaptation Evolution Strategy |
| **CS** | Cuckoo Search |
| **DR** | Discrepancy Ratio |
| **ELM** | Extreme Learning Machine |
| **EPR** | Evolutionary Polynomial Regression |
| **FS** | Feature selection |
| **FBNN** | Feedforward Backpropagation Neural Network |
| **FFA** | Firefly Algorithm |
| **GA** | Genetic Algorithm |
| **GEP** | Gene Expression Programming |
| **GMDH** | Group Method of Data Handling |
| **GA** | Genetic Algorithm |
| **GP** | Genetic Programming |
| **GEP** | Genetic Expression Programming |
| **GPR** | Gaussian Processes Regression |
| **GPR-FS** | Gaussian Processes Regression with Feature Selection |
| **GRC** | Granular Computing |
| **HHO** | Harris Hawks Optimization |
| **ICA** | Imperialist Competitive Algorithm |
| **LM** | Levenberg–Marquardt |
| **LSSVM** | least square support vector machine |
| **M5P** | M5 Model Tree |
| **MAD** | Mean Absolute Deviation |
| **MARS** | Multivariate Adaptive Regression Splines |
| **MGP** | Multigene Genetic Programming |
| **ML** | Machine Learning |
| **MLR** | Multiple Linear Regression |
| **MGP** | Multigene Genetic Programming |
| **NF-GMDH** | Neuro-Fuzzy-based Group Method of Data Handling |
| **PCA** | principal component analysis |
| **POMGGP** | Pareto-Optimal-Multigene Genetic Programming |
| **PSO** | Particle Swarm Optimization |
| **RF** | Random Forests |
| **RBFNN** | Radial Basis Function Neural Network |
| **RMD** | Root Mean Discrepancy |
| **RMSE** | Root Mean Squared Error |
| **SSMD** | Selection of Maximum Dissimilarity Method |
| **SVR** | Support Vector Regression |
| **SVR-FS** | Support Vector Regression with Feature Selection |
WOA Whale Optimization Algorithm

List of abbreviations

- $B$ Channel width
- $H$ Cross-sectional average flow depth
- $U$ Cross-sectional average flow velocity
- $u^*$ Shear velocity
- $U/u^*$ Relative shear velocity
- $Q$ Flow discharge
- $\beta$ Channel shape parameter
- $\sigma$ Channel sinuosity
- $K_X$ Longitudinal dispersion coefficient

Data and source availability

The datasets generated during and/or analysed during the current study are available in the github repository, https://github.com/LGoliatt/hybrid-evolutionary-ml-longitudinal-dispersion.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Alizadeh M., Shabani A., & Kavianpour M. (2017). Predicting longitudinal dispersion coefficient using ANN with metaheuristic training algorithms. *International Journal of Environmental Science and Technology*, 14(11), 2399–2410. https://doi.org/10.1007/s13762-017-1307-1

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. https://doi.org/10.1007/s11269-017-1611-z

Alizadeh M. J., Shahheydari H., Kavianpour M. R., Shamloo H., & Barati R. (2017). Prediction of longitudinal dispersion coefficient in natural rivers using a cluster-based Bayesian network. *Environmental Earth Sciences*, 76(2), 86. https://doi.org/10.1007/s12665-016-6379-6

Altunkaynak A. (2016). Prediction of longitudinal dispersion coefficient in natural streams by prediction map. *Journal of Hydro-Environment Research*, 12, 105–116. https://doi.org/10.1016/j.jher.2016.05.001

Azmathulla H. M., & Wu F. C. (2011). Support vector machine approach for longitudinal dispersion coefficients in natural streams. *Applied Soft Computing*, 11(2), 2902–2905. https://doi.org/10.1016/j.asoc.2010.11.026

Azar N. A., Milan S. G., & Kayhomayoon Z. (2021). The prediction of longitudinal dispersion coefficient in natural streams using LS-SVM and ANFIS optimized by Harris hawk optimization algorithm. *Journal of Contaminant Hydrology*, 240, 103781. https://doi.org/10.1016/j.jconhyd.2021.103781

Baek K. O. (2019). Deriving longitudinal dispersion coefficient based on Shiono and Knight model in open channel. *Journal of Hydraulic Engineering*, 145(3), 06019001. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001570

Balf M. R., Noori R., Berndtsson R., Ghaemi A., & Ghiassi B. (2018). Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers. *Journal of Water Supply: Research and Technology-Aqua*, 67(5), 447–457. https://doi.org/10.2166/aqua.2018.021

Belyaev M., Burnaev E., & Kapushev Y. (2014). Exact inference for Gaussian process regression in case of big data with the Cartesian product structure. Preprint. arXiv:1403.6573.

Beyer H. G., & Schwefel H. P. (2002, March 1). Evolution strategies – a comprehensive introduction. *Natural Computing*, 1(1), 3–52. https://doi.org/10.1023/A:1015059928466

Cai J., Luo J., Wang S., & Yang S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Camacho Suarez V., Schellart A., Brevis W., & Shucksmith J. (2019). Quantifying the impact of uncertainty within the longitudinal dispersion coefficient on concentration dynamics and regulatory compliance in rivers. *Water Resources Research*, 55(5), 4393–4409. https://doi.org/10.1029/2018WR023417

Camporeale C., & Ridolfi L. (2010). Interplay among river meandering, discharge stochasticity and riparian vegetation. *Journal of Hydrology*, 382(1), 138–144. https://doi.org/10.1016/j.jhydrol.2009.12.024

Carter H. H., & Okubo A. (1972). Longitudinal dispersion in nonuniform flow. *Water Resources Research*, 8(3), 648–660. https://doi.org/10.1029/WR008i003p00048

Chaudhary A., & Hantush M. M. (2017). Bayesian Monte Carlo and maximum likelihood approach for uncertainty estimation and risk management: Application to lake oxygen recovery model. *Water Research*, 108, 301–311. https://doi.org/10.1016/j.watres.2016.11.012

Chen K. Y. (2007). Forecasting systems reliability based on support vector regression with genetic algorithms. *Reliability Engineering & System Safety*, 92(4), 423–432. https://doi.org/10.1016/j.ress.2005.12.014

Creaco E., Berardi L., Sun S., Giustolisi O., & Savic D. (2016). Improvement on the existing equations for predicting longitudinal dispersion coefficient using field experimental data and 1D numerical modeling by multiobjective evolutionary polynomial regression paradigm. *Water Resources Research*, 52(4), 2403–2419. https://doi.org/10.1002/2015WR017971

Data C. U. F. E. (2019). Estimation of longitudinal dispersion coefficient using field experimental data and 1D numerical model of solute transport. In *Advanced Technologies, Systems, and Applications IV-Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies* (Vol. 83, p. 305).

Dehghani M., Zargar M., Riahi-Madvar H., & Memarzadeh R. (2018). Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers. *Journal of Hydraulic Engineering*, 128(10), 901–916. https://doi.org/10.1061/(ASCE)0733-9429(2002)128:10(901)

Deng Z. Q., Singh V. P., & Bengtsson L. (2001). Longitudinal dispersion coefficient in single-channel streams. *Journal of Hydraulic Engineering*, 127(11), 919–927. https://doi.org/10.1061/(ASCE)0733-9429(2001)127:11(919)
Desiante W. L., Minas N. S., & Fenner K. (2021). Micropollutant biotransformation and bioaccumulation in natural stream biofilms. *Water Research*, 193, 116846. https://doi.org/10.1016/j.watres.2021.116846

Disley T., Gharabaghi B., Mahboubi A., & McBean E. (2015). Predictive equation for longitudinal dispersion coefficient. *Hydrological Processes*, 29(2), 161–172. https://doi.org/10.1002/hyp.v29.2

Elder J. (2018). The apparent paradox of complexity in ensemble modeling. In N. Risbet, G. Miner, & K. Yale (Eds.), *Handbook of statistical analysis and data mining applications* (2nd ed., pp. 705–718). Academic Press.

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

Farzadkhoo M., Keshavarzi A., Hamidifar H., & Ball J. (2019, October). Flow and longitudinal dispersion in channel with partly rigid floodplain vegetation. In *Proceedings of the Institution of Civil Engineers-Water Management* (Vol. 172, No. 5, pp. 229–240). Thomas Telford Ltd.

Farzadkhoo M., Keshavarzi A., Hamidifar H., & Javan M. (2018). A comparative study of longitudinal dispersion models in rigid vegetated compound meandering channels. *Journal of Environmental Management, 217*, 78–89. https://doi.org/10.1016/j.jenvman.2018.03.084

Fuat Toprak Z., & Savci M. E. (2007). Longitudinal dispersion coefficient modeling in natural channels using fuzzy logic. *CLEAN–Soil, Air, Water, 35*(6), 626–637. https://doi.org/10.1002/ISSN1863-0650

Ghaemi A., Zbian T., Pirzadeh B., Monfared S. H., & Mosavi A. (2021). Reliability-based design and implementation of crow search algorithm for longitudinal dispersion coefficient estimation in rivers. *Environmental Science and Pollution Research, 28*(2), 35971–35990. https://doi.org/10.1007/s11356-021-12651-0

Ghiasi B., Jodeiri A., & Andik B. (2021). Using a deep convolutional network to predict the longitudinal dispersion coefficient. *Journal of Contaminant Hydrology, 240*, 103798. https://doi.org/10.1016/j.jconhyd.2021.103798

Ghiasi B., Sheikhzian H., Zeynolabedin A., & Niksokhan M. H. (2019). Granular computing – neural network model for prediction of longitudinal dispersion coefficients in rivers. *Water Science and Technology, 80*(10), 1880–1892. https://doi.org/10.2166/wst.2020.006

Gunn S. R. (1998). Support vector machines for classification and regression (ISIS Technical Report 14).

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

Hamidifar H., Omid M. H., & Keshavarzi A. (2015). Longitudinal dispersion in waterways with vegetated floodplain. *Ecological Engineering, 84*, 398–407. https://doi.org/10.1016/j.ecoleng.2015.09.048

Hansen N. (2006). The CMA evolution strategy: A comparing review. In J. A. Lozano, P. Larrañaga, I. Inza, & E. Bengoetxea (Eds.), *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms* (pp. 75–102). Springer.

Hansen N., & Ostermeier A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation, 9*(2), 159–195. https://doi.org/10.1162/106365601750 190398

Jimenez F., Sanchez G., Garcia J., Sciavicco G., & Miralles L. (2017). Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing, 234*, 75–92. https://doi.org/10.1016/j.neucom.2016.12.045

Kargar K., Samadianfard S., Parsa J., Nabipour N., Shamshirband S., Mosavi A., & Chau K. W. (2020). Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics, 14*(1), 311–322. https://doi.org/10.1080/19942060.2020.1712260

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

Kotu V., & Deshpande B. (2019). Data science process. In V. Kotu & B. Deshpande (Eds.), *Data science* (2nd ed., pp. 19–37). Morgan Kaufmann.

Kumar R. (2020). The generalized modified Bessel function and its connection with Voigt line profile and Humbert functions. *Advances in Applied Mathematics, 114*, 101986. https://doi.org/10.1016/j.aam.2019.101986

Li Z., Lin X., Zhang Q., & Liu H. (2020). Evolution strategies for continuous optimization: A survey of the state-of-the-art. *Swarm and Evolutionary Computation, 56*, 100694. https://doi.org/10.1016/j.swevo.2020.100694

Li Z., Zhang Q., Lin X., & Zhen H. (2018). Fast covariance matrix adaptation for large-scale black-box optimization. *IEEE Transactions on Cybernetics, 50*(5), 2073–2083. https://doi.org/10.1109/TCYB.6221036

Liu S., Wang X., Liu M., & Zhu J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics, 1*(1), 48–56. https://doi.org/10.1007/j. visinf.2017.01.006

Memarzadeh R., Zadeh H. G., Dehghani M., Riahi-Madvar H., Seifi A., & Mortazavi S. M. (2020). A novel equation for longitudinal dispersion coefficient prediction based on the hybrid of SSMD and whale optimization algorithm. *Science of The Total Environment, 716*, 137007. https://doi.org/10.1016/j.scitotenv.2020.137007

Meng J., & Li H. (2017). An efficient stochastic approach for flow in porous media via sparse polynomial chaos expansion constructed by feature selection. *Advances in Water Resources, 105*, 13–28. https://doi.org/10.1016/j.advwatres.2017.04.019

Milišić H., Hadžić E., & Jusić S. (2019, June). Estimation of longitudinal dispersion coefficient using field experimental data and 1D numerical model of solute transport. In *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies* (pp. 305–323). Springer.

Mirzaei A., Mohsenzadeh Y., & Sheikhzadeh H. (2017). Variational relevant sample-feature machine: A fully Bayesian approach for embedded feature selection. *Neurocomputing, 241*, 181–190. https://doi.org/10.1016/j.neucom.2017.02.057

Moreno-Rodenas A. M., Tscheikner-Gratl F., Langeveld J. G., & Clemens F. H. (2019). Uncertainty analysis in a large-scale water quality integrated catchment modelling study. *Water...
Noori R., Karbassi A., Farokhnia A., & Dehghani M. (2009). Evaluation of neuro-fuzzy GMHD-based particle swarm optimization to predict longitudinal dispersion coefficient in rivers. *Environmental Earth Sciences*, 57(2), 157. https://doi.org/10.1007/s12665-015-4877-6

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. https://doi.org/10.1007/s40710-018-0334-3

Noori R., Deng Z., Kiaghadi A., & Kachoosangi F. T. (2016). How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers? *Journal of Hydraulic Engineering*, 142(1), 04015039. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001062

Noori R., Karbassi A., Farokhnia A., & Dehghani M. (2009). Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. *Environmental Engineering Science*, 26(10), 1503–1510. https://doi.org/10.1089/ees.2008.0360

Noori R., Karbassi A., Mehdizadeh H., Vesali-Naseh M., & Sabahi M. (2011). A framework development for predicting the longitudinal dispersion coefficient in natural streams using an artificial neural network. *Environmental Progress & Sustainable Energy*, 30(3), 439–449. https://doi.org/10.1002/ep.v30.3

Noori R., Kerachian R., Darban A. K., & Shakibaieina A. (2007). Assessment of importance of water quality monitoring stations using principal components analysis and factor analysis: A case study of the Karoon river. *Journal of Water and Wastewater*, 63(3), 60–69.

Noori R., Mirchi A., Hooshyaripor F., Bhattarai R., Haghighi A. T., & Klove B. (2021). Reliability of functional forms for calculation of longitudinal dispersion coefficient in rivers. *Science of The Total Environment*, 791, 148394. https://doi.org/10.1016/j.scitotenv.2021.148394

Park H., & Hwang J. H. (2019). Quantification of vegetation arrangement and its effects on longitudinal dispersion in a channel. *Water Resources Research*, 55(5), 4488–4498. https://doi.org/10.1029/2019WR024807

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. https://doi.org/10.1504/IJHST.2018.095537

Paulson R. W. (1970). Variation of the longitudinal dispersion coefficient in the Delaware River Estuary as a function of freshwater inflow. *Water Resources Research*, 6(2), 516–526. https://doi.org/10.1029/WR006i002p00516

Perucca E., Camporeale C., & Ridolfi L. (2009). Estimation of the dispersion coefficient in rivers with riparian vegetation. *Advances in Water Resources*, 32(1), 78–87. https://doi.org/10.1016/j.adwres.2008.10.007

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. https://doi.org/10.1016/j.eswa.2008.10.043

Riahi-Madvar H., Dehghani M., Parmar K. S., Nabipour N., & Shamshirband S. (2020). Improvements in the explicit estimation of pollutant dispersion coefficient in rivers by subset selection of maximum dissimilarity hybridized with ANFIS-firefly algorithm (FFA). *IEEE Access*, 8, 60314–60337. https://doi.org/10.1109/ACCESS.2020.2979927

Riahi-Madvar H., Dehghani M., Seifi A., & Singh V. P. (2019). Pareto optimal multigene genetic programming for prediction of longitudinal dispersion coefficient. *Water Resources Management*, 33(3), 905–921. https://doi.org/10.1007/s11269-018-2139-6

Ros R., & Hansen N. (2008). *A simple modification in CMA-ES achieving linear time and space complexity* (Research Report No. RR-6498). INRIA. https://hal.inria.fr/inria-00270901

Rutherford J. C. (1994). *River mixing*. John Wiley & Sons.

Saberi-Movahed F., Najafzadeh M., & Mehrpooya A. (2020). Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: Training group method of data handling using extreme learning machine conceptions. *Water Resources Management*, 34(2), 529–561. https://doi.org/10.1007/s11269-019-02463-w

Sahay R. R. (2013). Predicting longitudinal dispersion coefficients in sinuous rivers by genetic algorithm. *Journal of Hydrology and Hydromechanics*, 61(3), 214–221. https://doi.org/10.2478/johh-2013-0028

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

Sahin S. (2014). An empirical approach for determining longitudinal dispersion coefficients in rivers. *Environmental Processes*, 1(3), 277–285. https://doi.org/10.1007/s40710-014-0018-6

Sattar A. M., & Gharabaghi B. (2015). Gene expression models for prediction of longitudinal dispersion coefficient in streams. *Journal of Hydrology*, 524, 587–596. https://doi.org/10.1016/j.jhydrol.2015.03.016

Savickis J., Bottacin-Busolin A., Zaramella M., Sabokrouhiyeh N., & Marion A. (2016). Effect of a meandering channel on wetland performance. *Journal of Hydrology*, 535, 204–210. https://doi.org/10.1016/j.jhydrol.2016.01.082

Schulz E., Speekenbrink M., & Krause A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85, 1–16. https://doi.org/10.1016/j.jmp.2018.03.001

Seo I. W., & Baek K. O. (2004). Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams. *Journal of Hydrographic Engineering*, 130(3), 227–236. https://doi.org/10.1061/(ASCE)0733-9429(2004)130:3(227)

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

Sharafati A., Haghbin M., Motta D., & Yaseen Z. M. (2021). The application of soft computing models and empirical formulations for hydraulic structure scouring depth simulation: A comprehensive review, assessment and possible future research direction. *Archives of Computational Methods in Engineering*, 28(2), 423–447. https://doi.org/10.1007/s11831-019-09382-4

Shen C., Niu J., Anderson E. J., & Phanikumar M. S. (2010). Estimating longitudinal dispersion in rivers using acoustic doppler current profilers. *Advances in Water Resources*, 33(5), 535–544. https://doi.org/10.1016/j.adwres.2010.01.008
Shieh H. Y., Chen J. S., Lin C. N., Wang W. K., & Liu C. W. (2010). Development of an artificial neural network model for determination of longitudinal and transverse dispersivities in a convergent flow tracer test. Journal of Hydrology, 391(3–4), 367–376. https://doi.org/10.1016/j.jhydrol.2010.07.041

Shin J., Seo I. W., & Baek D. (2019). Longitudinal and transverse dispersion coefficients of 2D contaminant transport model for mixing analysis in open channels. Journal of Hydrology, 583, 124302. https://doi.org/10.1016/j.jhydrol.2019.124302

Shucksmith J., Boxall J., & Guymer I. (2011). Determining longitudinal dispersion coefficients for submerged vegetated flow. Water Resources Research, 47(10). https://doi.org/10.1029/2011WR010547

Steele J. (1991). Measurement of the longitudinal dispersion coefficient near a junction in a tidal river. Water Resources Research, 27(5), 839–844. https://doi.org/10.1029/91WR00115

Stein M. L. (1999). Interpolation of spatial data: Some theory for kriging (1st ed.). Springer.

Swamee P. K., Pathak S. K., & Sohrib M. (2000). Empirical relations for longitudinal dispersion in streams. Journal of Environmental Engineering, 126(11), 1056–1062. https://doi.org/10.1061/(ASCE)0733-9372(2000)126:11(1056)

Tan P., Wang X., & Wang Y. (2020). Dimensionality reduction in evolutionary algorithms-based feature selection for motor imagery brain–computer interface. Swarm and Evolutionary Computation, 52, 100597. https://doi.org/10.1016/j.swevo.2019.100597

Tao H., Al-Bedwry N. K., Khedher K. M., Shahid S., & Yaseen Z. M. (2021). River water level prediction in coastal catchment using hybridized relevance vector machine model with improved grasshopper optimization. Journal of Hydrology, 598, 126477. https://doi.org/10.1016/j.jhydrol.2021.126477

Tao H., Habib M., Aljarah I., Faris H., Afan H. A., & Yaseen Z. M. (2021). An intelligent evolutionary extreme gradient boosting algorithm development for modeling scour depths under submerged weir. Information Sciences, 570, 172–184. https://doi.org/10.1016/j.ins.2021.04.063

Tayfur G. (2006). Fuzzy, ANN, and regression models to predict longitudinal dispersion coefficient in natural streams. Hydrology Research, 37(2), 143–164. https://doi.org/10.2166/nh.2006.0012

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

Tenebe I. T., Ogbiye A. S., Omore D. O., & Emenike P. C. (2016). Estimation of longitudinal dispersion coefficient: A review. Cogent Engineering, 3(1), 1216244. https://doi.org/10.1080/23311916.2016.1216244

Timar G. (2003). Controls on channel sinuosity changes: A case study of the Tisza River, the Great Hungarian Plain. Quaternary Science Reviews, 22(20), 2199–2207 https://doi.org/10.1016/S0277-3791(03)00145-8 Fluvial response to rapid environmental change

Toprak Z. F., & Cigizoglu H. K. (2008). Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods. Hydrological Processes: An International Journal, 22(20), 4106–4129. https://doi.org/10.1002/hyp.v22:20

Toprak Z. F., Sen Z., & Savci M. E. (2004). Comment on ‘Longitudinal dispersion coefficients in natural channels’. Water Research, 38(13), 3139–3143. https://doi.org/10.1016/j.watres.2003.08.004

Tutmez B., & Yuceer M. (2013). Regression kriging analysis for longitudinal dispersion coefficient. Water Resources Management, 27(9), 3307–3318. https://doi.org/10.1007/s11269-013-0348-6

van Dijk W. M., Teske R., van de Lageweg W. I., & Kleinmans M. G. (2013). Effects of vegetation distribution on experimental river channel dynamics. Water Resources Research, 49(11), 7558–7574. https://doi.org/10.1002/2013WR013574

Vose D. (1996). Quantitative risk analysis: A guide to monte carlo simulation modelling. John Wiley & Sons.

Wang B., & Chen T. (2015). Gaussian process regression with improved grasshopper optimization. Journal of Hydrologic Engineering, 20(1), 1216244. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001196

Wang Y., & Huai W. (2016). Estimating the longitudinal dispersion coefficient in straight natural rivers. Journal of Hydraulic Engineering, 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. Journal of Hydrology, 544, 511–523. https://doi.org/10.1016/j.jhydrol.2016.11.058

Yaseen Z. M. (2021). An insight into machine learning methods era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. Chemosphere, 277, 130126. https://doi.org/10.1016/j.chemosphere.2021.130126

Yaseen Z. M., Sulaiman S. O., Deo R. C., & Chau K. W. (2019). An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. Journal of Hydrology, 569, 387–408. https://doi.org/10.1016/j.jhydrol.2018.11.069

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