Evaluation of the effective parameters on energy losses of rectangular and circular culverts via kernel-based approaches

Kiyoumars Roushangar, Ghazaleh Nasssaji Matin, Roghayeh Ghasempour and Seyed Mahdi Saghebian

ABSTRACT

Energy dissipation in culverts is a complex phenomenon due to the nonlinearity and uncertainties of the process. In the current study, the capability of Gaussian process regression (GPR) and support vector machine (SVM) as kernel-based approaches and the gene expression programming (GEP) method was assessed in predicting energy losses in culverts. Two types of bend loss in rectangular culverts and entrance loss in circular culverts with different inlet end treatments were considered. Various input combinations were developed and tested using experimental data. The OAT (one-at-a-time), factorial sensitivity analysis and Monte Carlo uncertainty analysis were used to select the effective parameters in modeling. The results of performance criteria proved the capability of the applied methods (i.e. high correlation coefficient (R) and determination coefficient (DC) and low root mean square error (RSME)). For rectangular culverts, the model with parameters *Fr* (Froude number) and θ (bend angle), and for circular culverts, the model with parameters *Fr* and *Hw/D* (depth ratio), were the superior models. It showed that using the bend downstream Froude number caused an increment in model efficiency. Among the four end inlet treatments, mitered flush to 1.5:1 fill slope inlet yielded more accurate prediction. The sensitivity and uncertainty analysis showed that θ and *Hw/D* had the most significant impact on modeling, and *Fr* had the highest uncertainty.

Key words | bend, culvert, energy loss, GPR, inlet end treatments, SVM

Kiyoumars Roushangar (corresponding author)
Ghazaleh Nasssaji Matin
Roghayeh Ghasempour
Department of Water Resource Engineering,
Faculty of Civil Engineering,
University of Tabriz,
Tabriz,
Iran
E-mail: *kroshangar@yahoo.com*

Kiyoumars Roushangar
Center of Excellence in Hydroinformatics,
University of Tabriz,
Tabriz,
Iran

Seyed Mahdi Saghebian
Department of Civil Engineering, Ahar Branch,
Islamic Azad University,
Ahar,
Iran

INTRODUCTION

A culvert is a hydraulically short segment of conduit which conveys stream flow through a roadway embankment or past some other type of flow obstruction. Numerous cross-sectional shapes are available. The most commonly used shapes include circular, box (rectangular), elliptical, pipe-arch, and arch. Prediction of the accurate amount of local loss in the culvert systems is important due to its impact on saving costs and time of construction processes and determination of the size, shape, and diameter of the culverts. In a culvert system, with decreasing energy loss, its effect on the upstream flow profile decreases. The energy loss is divided into two categories: major loss and minor (or local) loss, which in the culvert systems due to its short length, major energy loss is negligible compared with minor loss. In fact, major loss is caused due to the friction between the flow and pipe walls and since culverts are usually used in short lengths in practice, therefore, the frictional or longitudinal loss is negligible in comparison with minor loss. So far, various studies have been conducted to explain the complex phenomenon of energy losses in culvert systems with different geometries. Tullis (2012) investigated the minor loss in buried-invert culverts and determined the optimum section from the point of view of the least loss. Malone & Parr (2008) investigated bend loss in rectangular

doi: 10.2166/hydro.2019.221

culverts and proposed some graphs for calculating this parameter. Tullis *et al.* (2005) studied the inlet losses in elliptical culverts and offered a precise method for calculating the outlet loss of the culvert and to identify the best section with minimum loss. Anderson (2006) studied the worn-out culverts and embedding new culverts inside them and determined the local loss empirically. Kotowski *et al.* (2011) studied the inlet and outlet loss coefficient in the conduits and concluded that inlet loss coefficient in the pipes was not constant. However, due to the complexity and uncertainty of the local losses phenomenon, the results of the classical models are not general and under variable conditions do not present the same results. Therefore, it is essential to use other methods with more accuracy in predicting energy loss in culverts with different shapes under varied hydraulic conditions.

In recent years, the application of nonlinear machine learning (ML) (e.g. artificial neural networks (ANNs), neuro-fuzzy models (NF), genetic programming (GP), gene expression programming (GEP), support vector machine (SVM), and Gaussian process regression (GPR)) in water resources engineering has become viable leading to numerous publications in this field. A complete review of all the applications is beyond the scope of this paper and only some studies are mentioned here, such as assessing tree-based methods concepts, uses and limitations (Carvalho *et al.* 2018), modeling historical land use changes using ANN (Tayyebi *et al.* 2017), prediction of groundwater levels using data-driven models (Huang *et al.* 2017; Amaranto *et al.* 2018), estimation of hydraulic jump energy dissipation in channels with rough elements using SVM (Roushangar & Ghasempour 2018), prediction of pile scour using ANN and kernel methods (Ghazanfari-Hashemi *et al.* 2011; Pal *et al.* 2014), computing longitudinal dispersion coefficients in natural streams using SVM (Azamathulla & Wu 2011), real time hydrologic forecasting using EC-SVM (Yu *et al.* 2004), quantify runoff contributions from different land uses in tropical urban environments using GP (Meshgi *et al.* 2015), side weir discharge coefficient using SVM (Azamathulla *et al.* 2017), and forecasting monthly and seasonal streamflow using mixture-kernel GPR approach (Zhu *et al.* 2018).

In artificial intelligence models we are looking for a learning machine capable of finding an accurate approximation of a natural phenomenon, as well as expressing it in the form of an interpretable equation. However, this bias towards interpretability creates several new issues. The computer-generated hypotheses should take advantage of the already existing body of knowledge about the domain in question. However, the method by which we express our knowledge and make it available to a learning machine remains rather unclear (Babovic 2009). Machine learning, a branch of artificial intelligence, deals with the representation and generalization using a data learning technique. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory. There is a wide variety of machine learning tasks and successful applications (Mitchell 1997). In general, the task of a machine learning algorithm can be described as follows: Given a set of input variables and the associated output variable(s), the objective is learning a functional relationship for the input–output variables set. It should be noted that artificial intelligence models typically do not really represent the physics of a modeled process; they are just devices used to capture relationships between the relevant input and output variables. However, when the interrelationships among the relevant variables are poorly understood, finding the size and shape of the ultimate solution is difficult, and conventional mathematical analysis methods do not (or cannot) provide analytical solutions; these methods can predict the interest variable with more accuracy.

Due to the complexity and uncertainties of the energy losses process, the existing regression models do not show desired accuracy and their output is often associated with large errors. Therefore, the present research proposed kernel based models to predict the energy losses coefficient and also to investigate the best input models and determine the effective parameters of different shaped culverts. To the best of the authors' knowledge there is a lack of research on the comprehensive study of predicting local losses in culverts using artificial intelligence. In all previous studies the local loss coefficient in the culvert was measured and recorded experimentally at various velocities, but the relationship between this coefficient and the Froude and Reynolds numbers, and the geometric parameters, and the

dependency rate of this coefficient on these parameters, was not investigated. Therefore, this study aimed to assess the capability of GPR and SVM as kernel-based approaches for modeling the losses of culverts with different geometries. Also, the GEP method was used to develop new equations for predicting the local loss coefficient in different shaped culverts. In order to determine the most effective combination for modeling the losses of culverts, different input combinations were considered under two scenarios (losses due to the culvert bend and culvert entrance) and the impact of hydraulic characteristics and culvert shapes was assessed. In addition, the most important parameters in predicting the energy losses were determined using one-at-a-time (OAT) and factorial sensitivity analysis and Monte Carlo uncertainty sensitivity.

MATERIALS AND METHODS

The data sets

The data sets of laboratory experiments of local losses of culverts performed by Malone & Parr (2008) and Tullis (2012) were used in the present study. Malone & Parr (2008) studied bend losses in rectangular culverts. Laboratory experiments were performed in rectangular channels with abrupt bends. Bend angles of approximately 30, 45, 60, 75 and 90° were tested. Tullis (2012) conducted experiments on circular culverts in order to determine the entrance loss coefficient and the inlet control head discharge relationships for circular culverts with invert burial depths of 20, 40, and 50%. All buried-invert culverts were tested with four different end treatments: (1) thin-wall projecting, (2) mitered flush to 1.5:1.0 (horizontal to vertical) fill slope, (3) square-edged inlet with vertical headwall, and (4) 45° beveled entrance with vertical headwall. The ranges of some parameters used in the tests are given in Table 1 in which *Ke*, θ and *Fr*, and *Re* represent the entrance loss coefficient, culvert bend angle, Froude number, and Reynolds number respectively.

Kernel-based approaches

Kernel based approaches, such as GPR and SVM, are a relatively new important method based on the different kernel types which are based on initiation of statistical learning theory. Such models are capable of adapting themselves to predict any variable of interest via sufficient inputs. These methods can model non-linear decision boundaries, and there are many kernels to choose from. They are also fairly robust against overfitting, especially in high-dimensional space. However, the appropriate selection of kernel type is the most important step in the GPR and SVM due to its direct impact on the training and classification precision. In fact, these methods are memory intensive, trickier to tune due to the importance of picking the right kernel, and do not scale well to larger data sets. In these models we will be able to predict the proper behavior of the system, although we will not be able to characterize its intrinsic structure and behavior. In other words, we will be able to say what the model does, but not how. In addition to this, we will not be able to guarantee the behavior of such a model in regions not covered by the data from which the model was

Table 1 | Detail of various parameters from laboratory experiments used in this study

Bend loss			Entrance loss			
			Circular culvert (Tullis 2012)			
Rectangular culvert (Malone & Parr 2008)			Thin-wall projecting	Mitered to flush 1.5 h:1 v fill slope	Square edge inlet with vertical headwalls	45° beveled inlet with vertical headwalls
Parameters	Ke	0.157–1.078	0.157–1.03	0.42–0.93	0.3–0.6	0.22–0.38
	θ (radian)	0.523–1.578	–	–	–	–
	Fr	0.181–0.86	0.0124–1.058	0.01–0.81	0.43–0.97	0.049–1.05
	Re	42,138–140,590	14,408–30,711	18,743–268,0463	9,616–305,469	79,175–292,240
No. of data		190	66	65	45	48

constructed. This is due to the fact that the model covers only the relationships found within the given data (Babovic 2009).

Gaussian process regression

GPR models are based on the assumption that adjacent observations should convey information about each other. Gaussian processes are a way of specifying a priori directly over function space. This is a natural generalization of the Gaussian distribution, whose mean and covariance are a vector and matrix, respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions. Thus, due to prior knowledge about the data and functional dependencies, no validation process is required for generalization and GP regression models are able to understand the predictive distribution corresponding to the test input (Rasmussen & Williams 2006). A GP is defined as a collection of random variables, any finite number of which has a joint multivariate Gaussian distribution. Let $\chi \times \gamma$ represent the domains of inputs and outputs, respectively, from which n pairs (x_i, y_i) are drawn independently and identically distributed. For regression, assume that $y \subseteq \Re$; then, a GP on χ is defined by a mean function $\mu : \chi \rightarrow \Re$ and a covariance function $k : \chi \times \chi \rightarrow \Re$.

The main assumption of GP regression is that y is given by $y = f(x) + \xi$, where $\xi \sim \mathrm{N}(0, \sigma^2)$. In GP regression, for every input x there is an associated random variable $f(x)$, which is the value of the stochastic function f at that location. In this work, it is assumed that the observational error ξ is normally independent and identically distributed, with a mean value of zero ($\mu(x) = 0$), a variance of σ^2 and $f(x)$ drawn from the Gaussian process on χ specified by k. That is, $Y = (y_1, \ldots\ldots, y_n) \sim \mathrm{N}(0, K + \sigma^2 I)$ where $K_{ij} = k(x_i, x_j)$, and I is the identity matrix. Because $Y/X \sim \mathrm{N}(0, K + \sigma^2 I)$ is normal, so is the conditional distribution of test labels given the training and test data of $p(Y*/Y, X, X*)$. Then, one has $Y*/Y, X, X* \sim \mathrm{N}(\mu, \Sigma)$, where:

$$\mu = K(X*, X)(K(X, X) + \sigma^2 I)^{-1} Y \quad (1)$$

$$\Sigma = K(X*, X*) - \sigma^2 I - K(X*, X)(K(X, X) + \sigma^2 I))^{-1} K(X, X*) \quad (2)$$

If there are n training data and $n*$test data, then $K(X, X*)$ represents the $n \times n*$ matrix of covariances evaluated at all pairs of training and test data sets, and this is similarly true for the other values of $K(X, X)$, $K(X*, X)$ and $K(X*, X*)$; here X and Y are the vectors of the training data and training data labels y_i, whereas $X*$ is the vector of the test data. A specified covariance function is required to generate a positive semi-definite covariance matrix K, where $K_{ij} = k(x_i, x_j)$. The term of the kernel function used in SVM is equivalent to the covariance function used in GP regression. With the known values of kernel k and degree of noise σ^2, Equations (1) and (2) would be enough for inference. During the training process of GP regression models, one needs to choose a suitable covariance function as well as its parameters. In the case of GP regression with a fixed value of Gaussian noise, a GP model can be trained by applying Bayesian inference, i.e. maximizing the marginal likelihood. This leads to the minimization of the negative log-posterior:

$$p(\sigma^2, k) = \frac{1}{2} \mathrm{y}^T (K + \sigma^2 I)^{-1} \mathrm{y} + \frac{1}{2} \log |\mathrm{K} + \sigma^2 I| - \log p(\sigma^2) - \log p(k) \quad (3)$$

To find the hyperparameters, the partial derivative of Equation (3) can be obtained with respect to σ^2 and k, and minimization can be achieved by gradient descent. For more details about GP regression and different covariance functions, readers are referred to Kuss (2006). The optimal value of capacity constant (C) and the size of error-intensive zone (ε) in SVM and Gaussian noise in GPR are required due to their high impact on the accuracy of the mentioned regression approaches. The optimum values of these parameters were obtained after the trial-and-error process.

Support vector machine

Support vector machines as an intelligence approach are used in information categorization and data set classification. This approach, developed by Vapnik (1995), is known as structural risk minimization (SRM), which minimizes an upper bound on the expected risk, as opposed to

the traditional empirical risk (ERM) which minimizes the error on the training data. The SVM method is based on the concept of the optimal hyper plane that separates samples of two classes by considering the widest gap between two classes. SVR is an extension of SVM regression. The purpose of the SVR is to find a function having the most deviation from the actual target vectors for all given training data and to have it be as flat as possible (Smola 1996). Vapnik (1995) introduced the concept of kernel function for non-linear support vector regression. The most important step in the SVM is the appropriate selection of kernel type. In general, there are several types of kernel functions, namely linear, polynomial, radial basis function (RBF) and sigmoid functions. Due to the black-box nature of SVM and GPR models, the learned relationship between the inputs and output is not revealed. This requires cautious usage of the new model, such as GEP, and it should not be used beyond the ranges of the data for which it was trained.

Gene expression programming

Gene expression programming was developed by Ferreria (2001) using fundamental principles of genetic algorithms (GA) and genetic programming (GP). One strength of the GEP approach is that the creation of genetic diversity is extremely simplified as genetic operators work at the chromosome level. Another strength of GEP consists of its unique, multigenic nature, which allows the evolution of more complex programs composed of several subprograms. GEP as GA mimics the biological evolution to create a computer program for simulating a specified phenomenon. A GEP algorithm begins by selecting five elements such as the function set, terminal set, fitness function, control parameters, and stopping condition. There is a comparison between predicted values and actual values in each subsequent step. When desired results in accordance with previously selected error criteria are found, the GEP process is terminated. If the desired error criteria could not be found, some chromosomes are chosen by a method called roulette wheel sampling and they are mutated to obtain new chromosomes. After the desired fitness score is found, this process terminates and then the chromosomes are decoded for the best solution of the problem. The advantages of a system like GEP are clear from nature, but the most important are (Ferreria 2001): (1) the chromosomes are simple entities: linear, compact, relatively small, easy to manipulate genetically (replicate, mutate, recombine, etc.); (2) the expression trees are exclusively the expression of their respective chromosomes; they are entities upon which selection acts, and according to fitness, they are selected to reproduce with modification.

Performance criteria

In the current study, the model's performance was evaluated using three statistical parameters: correlation coefficient (R), determination coefficient (DC), and root mean square errors (RSME). It should be noted that the RMSE criteria has the same unit of the target parameter (K_e). However, science K_e is a non-dimensional parameter, therefore, RMSE is also a dimensional value. Expressions for performance criteria are as follows:

$$DC = 1 - \frac{\sum_{i=1}^{N}(l_o - l_p)^2}{\sum_{i=1}^{N}(l_o - \overline{l_p})^2},\ \mathrm{R} = \frac{\sum_{i=1}^{N}(l_o - \overline{l_o}) \times (l_p - \overline{l_p})}{\sqrt{\sum_{i=1}^{N}(l_o - \overline{l_o})^2 \times (l_p - \overline{l_p})^2}},$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(l_o - l_p)^2}{N}} \quad (4)$$

where l_o, l_p, $\overline{l_o}$, $\overline{l_p}$, N are, respectively: the measured values, predicted values, mean measured values, mean predicted values and number of data samples.

Simulation and models development

Input variables

The crucial step during the modeling process via an intelligent method is an appropriate selection of model input parameters. In this study, the local energy loss in culvert systems is expressed through a set of dimensionless variables. Based on Malone & Parr (2008) and Tullis (2012), the most important parameters in the energy loss (K_e) prediction in culverts with different shapes are as follows:

$V,\ D,\ g,\ \mu,\ \theta,\ \rho,\ Hw$

where V_1 is upstream flow velocity, μ is dynamic viscosity of water, g is acceleration due to gravity, ρ is density of water, D is culvert diameter, Hw is flow depth. Also, θ is a geometric parameter and indicates the bend angle along the culvert. From the dimensional analysis the above parameters can be expressed as follows:

$$Fr,\ \theta,\ Re,\ \frac{Hw}{D}$$

where $Fr = V/(g \times Hw)^{0.5}$ is the Froude number. The Froude number is a dimensionless value that describes different flow regimes of open channel flow. The Froude number is a ratio of inertial and gravitational forces. As flow passes through critical conditions its Froude number has a value of 1. For subcritical flow the depth is greater and the velocity lower, therefore the Froude number is always less than 1; for supercritical flow the opposite is true and the Froude number is always greater than 1. Calculation of the Froude number thus provides an immediate check on the type of flow and how near the flow conditions are to those at critical depth. $Re = VR/v$ is Reynolds number (R: hydraulic radius, v: kinematic viscosity). The Reynolds number is the ratio of inertial forces to viscous forces and is a convenient parameter for predicting if a flow condition will be laminar or turbulent. In this study, for developing the local loss models in rectangular and circular culverts, two scenarios with different input variables were considered (Figure 1). In the first scenario, the impact of culvert bend was evaluated and in the second scenario, the entrance energy loss

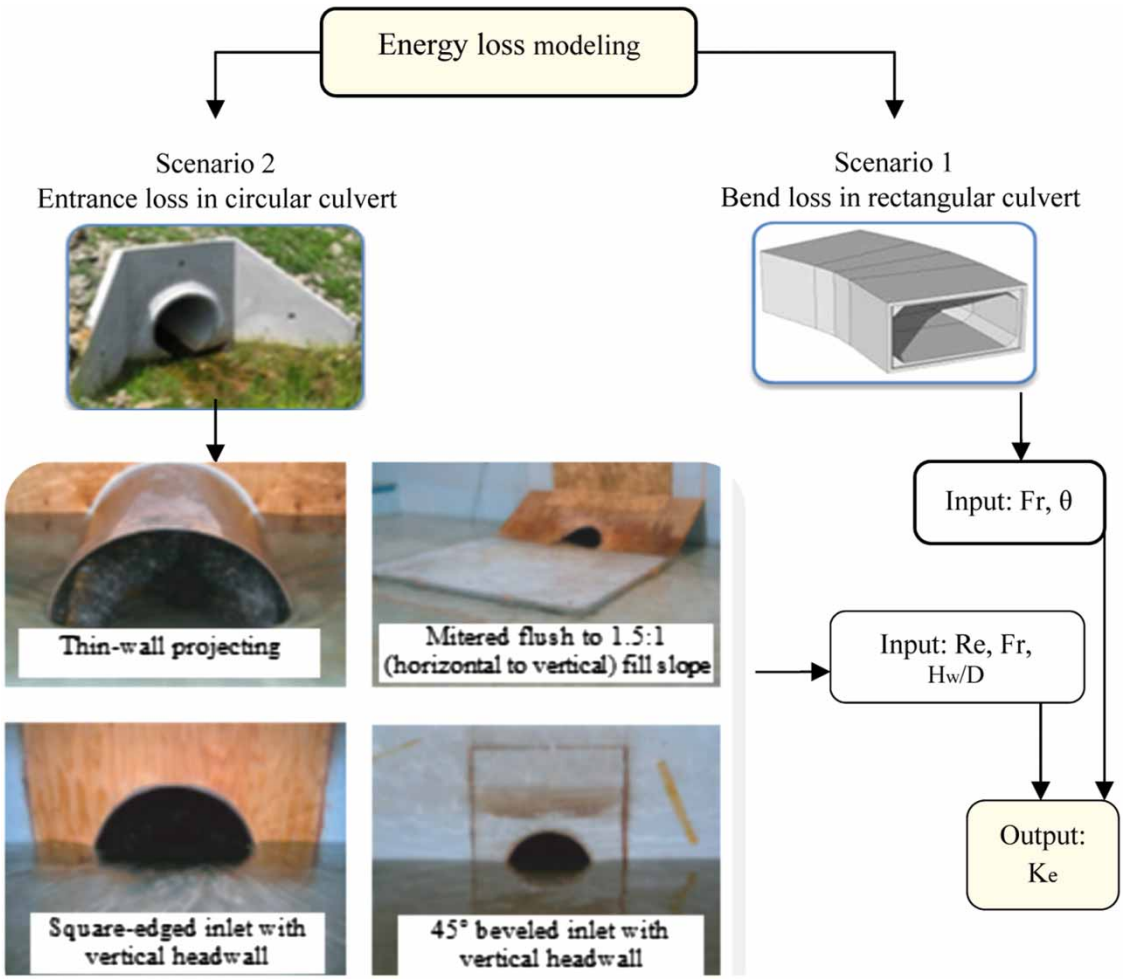

Figure 1 | Schematic view of different states considered in the study.

was investigated. Figure 1 illustrates the flowchart of considered scenarios. Developed models of GPR and SVM for predicting local losses are given in Table 2. In assessing the local loss in culvert bend, Froude numbers of three points (downstream and upstream of bend and at the bend) were used. It should be noted that for all models 75% of the data set was used for training goals and the remaining 25% of data were used for testing goals.

Table 2 | GPR, SVM, and GEP developed models

Bend energy loss		Entrance loss	
Model	**Input variables**	**Model**	**Input variables**
B(I)	Fr downstream	E(I)	Re
B(II)	Fr downstream, θ	E(II)	Re, Hw/D
B(III)	Fr average	E(III)	Fr
B(IV)	Fr average, θ	E(IV)	Fr, Hw/D
B (V)	Fr upstream		
B(VI)	Fr upstream, θ		

SVM, GPR, and GEP models development

The design of GP and SVM-based regression approaches involve the use of the concept of the kernel function. A number of kernels are discussed in the literature, but studies suggest a better performance by radial basis kernels for different civil engineering problems (Gill *et al.* 2006). In this study, for determining the best performance of SVM and GPR and selecting the best kernel function, the model B(II) from Scenario 1 in a rectangular culvert was predicted via SVM and GPR using various kernels. Figure 2 indicates the results of the statistical parameters of different kernels for this model. According to the results, using the kernel function of RBF in the SVM model led to better prediction accuracy in comparison

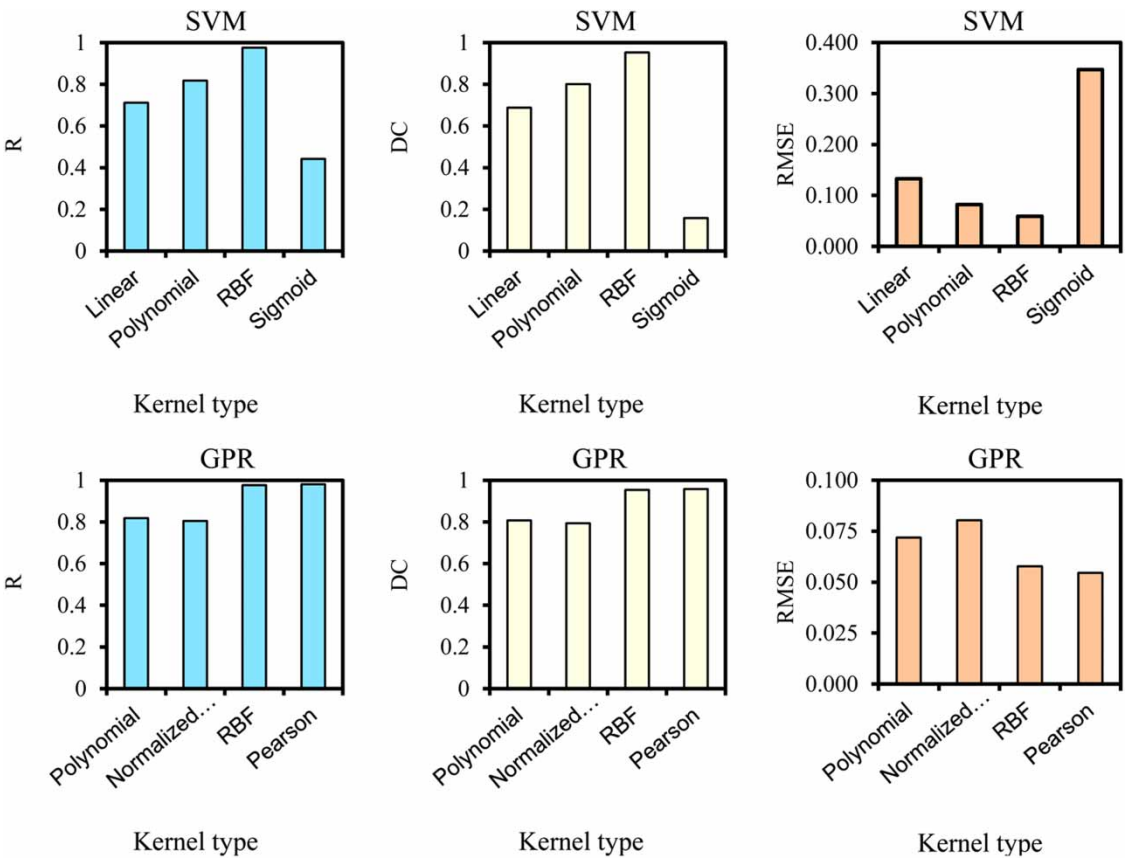

Figure 2 | Statistics parameters via SVM and GPR kernels function types for a testing set of model B(II) of rectangular culvert.

to the other kernels and for the GPR model using the kernel function of Pearson led to better prediction accuracy. Therefore, RBF and Pearson kernels were used as a core tool of SVM and GPR which were applied for the rest of the models.

GEP was trained for energy losses prediction in rectangular and circular culverts. Basic arithmetic operators of (+, –, *, /) and several mathematical functions (exp, X^2, X^3, $\sqrt{}$) were utilized as the GEP function set. The architecture of the chromosomes, including number of chromosomes (25-30-35), head size (7-8) and number of genes (3–4), were selected and different combinations of the mentioned parameters were tested. The model was run for a number of generations and was stopped when there was no significant change in the fitness function value and coefficient of correlation. It is observed that the model with the number of chromosomes of 30, head size of 7, and number of genes of 3 yielded better results. Also, addition and multiplication were tested as linking functions and it was found that linking the sub-ETs by addition represented better fitness values. One of the important steps in preparing the GEP model is to choose the set of genetic operators. In the current study, a combination of all genetic operators (recombination, mutation, transposition, and crossover) was used for this aim. Parameters of the optimized GEP model are shown in Table 3.

Table 3 | Optimized parameters of GEP models used in this study

Description of parameter	Setting of parameter	Description of parameter	Setting of parameter
Function set	+, –, ×, /, X^2, X^3, $\sqrt{}$	Fitness function error type	Root mean square error
Chromosomes	30	Mutation rate	0.044
Head size	7	Inversion, IS and RIS transposition rate	0.1
Number of genes	3	One- and two-point recombination rate	0.3
Linking function	Addition	Gene recombination and transposition rate	0.1

RESULTS AND DISCUSSION

Developed models for rectangular culvert with bend (Scenario 1)

For evaluating the impact of culvert bend on energy loss in culvert systems, several models were developed based on flow characteristic (in the term of Froude number at three points of the culvert) and bend angle. All of the SVM, GPR, and GEP models were trained and tested to carry out the local loss prediction in culverts. The obtained results are listed in Table 4 and shown in Figure 3. From the RMSE, R, and DC viewpoints (i.e. highest R and DC and lowest RMSE), it can be seen that the models with input parameters of Fr and θ show better performance than the models which only use Fr as an input variable. According to the results, among the six developed models, the model B(II) led to more accurate results. It could be inferred that adding the geometric θ parameter to the input combination caused an increment in models efficiency and this parameter had a significant impact on local loss prediction process. Also, the Froude number of the bend downstream led to more accurate outcomes. However, considering the results of the model B(VI), using the upstream Froude number did not lead to undesirable results. This issue confirms that the local loss can be estimated using upstream flow parameters when there is no information about hydraulic conditions of bend downstream. According to the results, it can be seen that the results of the GPR models are slightly more accurate than the SVM and GEP models. The comparison of observed and predicted local loss for the superior model is shown in Figure 3. The mathematical expression of GEP for the best model is as follows:

$$K_e = \frac{\sqrt{Fr}}{-4.66\theta} + (Fr^4 - \theta) \times \left(\frac{Fr}{0.996}\right)^3 + 0.875\theta \quad (5)$$

Developed models for circular culverts with different end inlet treatments (Scenario 2)

For Scenario 2, different models were developed based on flow condition and culvert diameters in order to assess the entrance loss in circular culverts with different end inlet treatments. The obtained results of GPR, SVM, and GEP

Table 4 | Statistical parameters of the GPR and SVM models; Scenario 1

Models		Train			Test		
		R	DC	RMSE	R	DC	RMSE
B(I)	SVM	0.601	0.503	0.243	0.564	0.487	0.271
	GPR	0.604	0.507	0.241	0.566	0.490	0.268
	GEP	0.521	0.402	0.302	0.505	0.334	0.312
B(II)	SVM	0.981	0.964	0.053	0.976	0.956	0.058
	GPR	0.985	0.971	0.051	0.981	0.961	0.055
	GEP	0.973	0.947	0.064	0.972	0.943	0.068
B(III)	SVM	0.701	0.612	0.241	0.659	0.537	0.266
	GPR	0.705	0.613	0.240	0.661	0.539	0.263
	GEP	0.691	0.532	0.260	0.618	0.438	0.298
B(IV)	SVM	0.980	0.961	0.056	0.976	0.951	0.062
	GPR	0.984	0.965	0.053	0.981	0.955	0.059
	GEP	0.972	0.945	0.065	0.971	0.942	0.075
B(V)	SVM	0.705	0.529	0.272	0.647	0.424	0.283
	GPR	0.709	0.530	0.270	0.650	0.427	0.280
	GEP	0.652	0.512	0.298	0.611	0.412	0.302
B(VI)	SVM	0.977	0.962	0.055	0.973	0.949	0.064
	GPR	0.982	0.964	0.056	0.978	0.953	0.062
	GEP	0.965	0.931	0.074	0.962	0.926	0.086

models are listed in Table 5 and shown in Figure 4. The superior performance for this state and for all end inlet treatments was obtained for the model E(IV), in which the inputs were Fr and Hw/D. According to Table 5, it seems that for modeling entrance loss in culverts, using relative flow depth as the input parameter improved the efficiency of the models. Comparing the models E(I) and E(III) and considering the RMSE values, the obtained error percentage for the model E(I) is almost 8–22% more than the model E(II), therefore, in models with only one input parameter, using the Fr number led to better prediction than using the Re number. Among the four end inlet treatments, culverts with a mitered flush to 1.5:1 (horizontal to vertical) fill slope yielded a more accurate prediction. The mathematical expressions of GEP for all cases are as follows.

Thin-wall projecting:

$$k_e = 1 - \left(0.01\frac{H_w}{D}\right)^3 + \frac{\left(0.73\left(\frac{H_w}{D}\right)^2\right)^3}{2.3 - \left(\frac{H_w}{D}\right)^3} - \frac{Fr^2}{3.4Fr^3 - \sqrt{Fr}} \quad (6)$$

Mitered flush to 1.5:1 (horizontal to vertical) fill slope:

$$k_e = -0.55\left(0.02Fr^2 + \frac{H_w}{D}\right) + \frac{\sqrt{\frac{H_w}{D}}}{\frac{\sqrt{Fr}}{\frac{H_w}{D} - Fr} + 5.82} + \sqrt{\sqrt{\frac{H_w}{D}}^3} \quad (7)$$

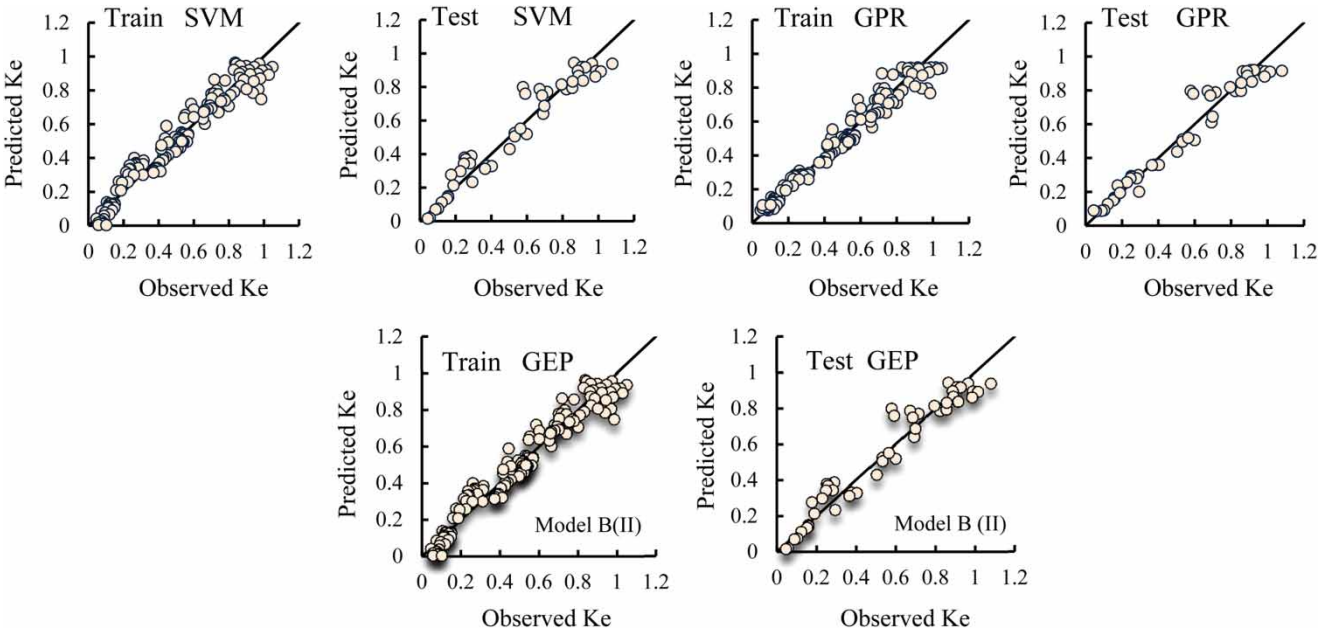

Figure 3 | Comparison of observed and predicted local loss for best models of Scenario 1; model B(II).

Table 5 | Statistical parameters of the GPR, SVM, and GEP models; Scenario 2

		Model					
		Train			Test		
		R	DC	RMSE	R	DC	RMSE
E(I)	SVM	0.705	0.632	0.075	0.634	0.547	0.073
	GPR	0.716	0.635	0.073	0.653	0.574	0.071
	GEP	0.622	0.62	0.088	0.617	0.592	0.092
E(II)	SVM	0.841	0.732	0.059	0.832	0.681	0.068
	GPR	0.842	0.733	0.057	0.837	0.684	0.065
	GEP	0.831	0.720	0.058	0.790	0.610	0.080
E(III)	SVM	0.820	0.692	0.065	0.695	0.612	0.071
	GPR	0.822	0.695	0.063	0.716	0.643	0.068
	GEP	0.815	0.659	0.068	0.593	0.63	0.073
E(IV)	SVM	0.846	0.769	0.055	0.832	0.741	0.058
	GPR	0.853	0.770	0.054	0.857	0.748	0.056
	GEP	0.851	0.750	0.060	0.830	0.725	0.063
E(I)	SVM	0.852	0.747	0.073	0.833	0.741	0.075
	GPR	0.856	0.751	0.070	0.837	0.745	0.073
	GEP	0.843	0.735	0.08	0.826	0.69	0.09
E(II)	SVM	0.980	0.960	0.033	0.975	0.943	0.041
	GPR	0.985	0.962	0.030	0.979	0.947	0.039
	GEP	0.964	0.924	0.046	0.962	0.88	0.057
E(III)	SVM	0.941	0.854	0.056	0.912	0.856	0.059
	GPR	0.944	0.857	0.054	0.918	0.859	0.056
	GEP	0.911	0.822	0.07	0.908	0.82	0.06
E(IV)	SVM	0.979	0.959	0.032	0.978	0.954	0.037
	GPR	0.984	0.963	0.029	0.979	0.959	0.035
	GEP	0.975	0.950	0.033	0.962	0.92	0.040
E(I)	SVM	0.820	0.685	0.041	0.812	0.635	0.052
	GPR	0.824	0.688	0.039	0.817	0.636	0.048
	GEP	0.816	0.641	0.039	0.790	0.560	0.064
E(II)	SVM	0.887	0.791	0.031	0.876	0.671	0.036
	GPR	0.891	0.795	0.029	0.883	0.673	0.034
	GEP	0.883	0.781	0.032	0.860	0.641	0.038
E(III)	SVM	0.834	0.752	0.033	0.817	0.662	0.046
	GPR	0.835	0.755	0.031	0.821	0.666	0.043
	GEP	0.820	0.661	0.036	0.810	0.592	0.058
E(IV)	SVM	0.905	0.817	0.028	0.894	0.742	0.034
	GPR	0.911	0.822	0.025	0.896	0.745	0.032
	GEP	0.889	0.791	0.030	0.876	0.732	0.034
E(I)	SVM	0.732	0.605	0.027	0.699	0.532	0.036
	GPR	0.736	0.608	0.026	0.702	0.533	0.034
	GEP	0.687	0.504	0.029	0.664	0.470	0.035
E(II)	SVM	0.821	0.672	0.021	0.812	0.632	0.024
	GPR	0.824	0.677	0.018	0.816	0.633	0.023
	GEP	0.753	0.630	0.022	0.680	0.612	0.024
E(III)	SVM	0.795	0.671	0.022	0.714	0.602	0.029
	GPR	0.799	0.673	0.020	0.718	0.605	0.028
	GEP	0.730	0.582	0.024	0.654	0.530	0.032
E(IV)	SVM	0.914	0.833	0.013	0.895	0.736	0.015
	GPR	0.919	0.837	0.011	0.899	0.739	0.014
	GEP	0.802	0.784	0.018	0.730	0.651	0.021

Square-edged inlet with vertical headwall:

$$K_e = \frac{7.6 + Fr}{438.97\left[\frac{9.36D}{H_w} + \left(\frac{H_w}{D} - 9.36\right)\right]} + \frac{Fr}{\left(\sqrt[4]{\frac{H_w}{D \times Fr}^{3}}\right) + Fr - 0.16\left(\frac{D \times Fr \times \left[0.77 - \frac{H_w}{D}\right]}{H_w}\right)} \quad (8)$$

45° beveled inlet with vertical headwall:

$$K_e = \frac{Fr}{\left(0.21Fr \times \frac{H_w}{D}\right) + (Fr - 8.14)} + (0.22 - 0.011Fr^3) + \left(\frac{1}{Fr}\right)^3 \quad (9)$$

Sensitivity and uncertainty analysis

Sensitivity analysis was used to evaluate the effect of different employed parameters on local loss prediction via GPR. There are different methods for carrying out sensitivity analysis, such as local or global, quantitative or qualitative or one-at-a-time (OAT), etc. One of the most simple and most common approaches is that of changing OAT, to see what effect this produces on the output. OAT sensitivity analysis essentially consists of selecting a base parameter setting (nominal set) and varying one parameter at a time while keeping all other parameters fixed (hence it is referred to as a local method). An important use of OAT is to reveal the form of the relationship between the varied parameter and the output, given that all other parameters have their nominal values (Holvoet *et al.* 2005). In this study, for evaluating the impact of each parameter, the model was run with all input parameters and then one of the input parameters was eliminated and the model was re-run. Based on the results from Figure 5, it could be deduced that in the bend loss prediction process the variable θ, and in entrance loss prediction process the variable Hw/D, had the most significant impact on local loss, respectively. With eliminating the θ and Hw/D variables the amount of RMSE error criteria increased to 0.212 and 0.066, respectively. Also, the uncertainty analysis was performed in order to determine the uncertainty of each parameter.

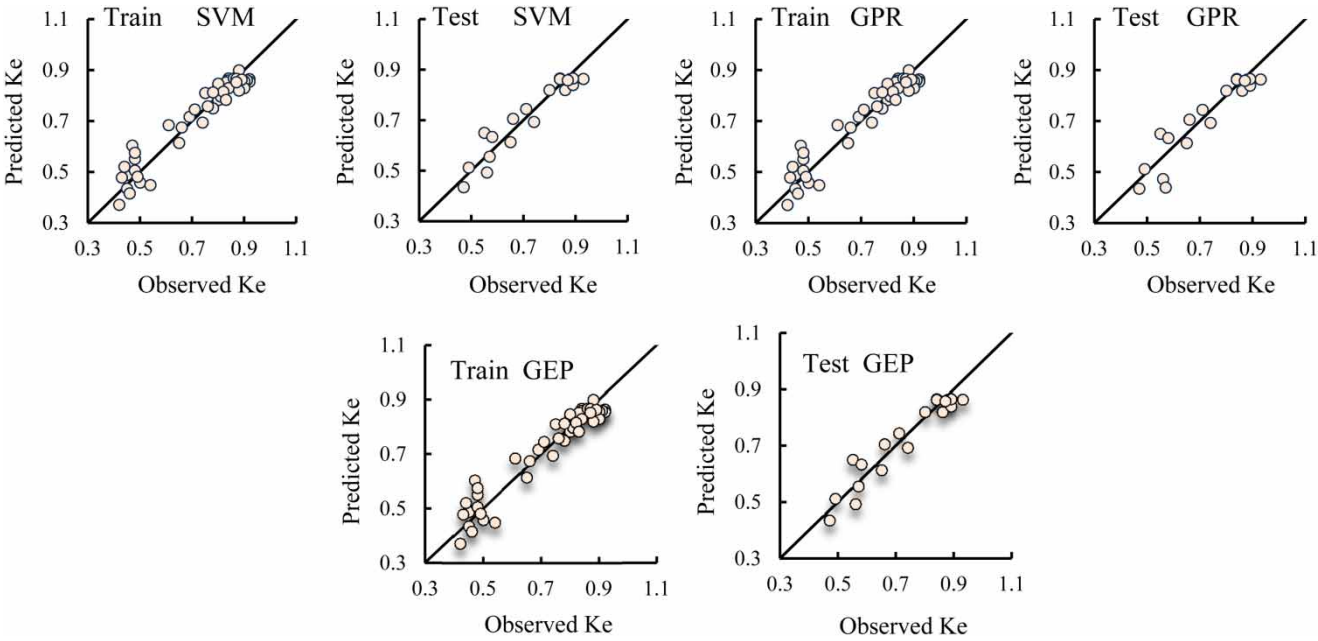

Figure 4 | Comparison of observed and predicted local loss for best models of Scenario 2 for circular culvert with mitered flush to 1.5:1 (horizontal to vertical) fill slope.

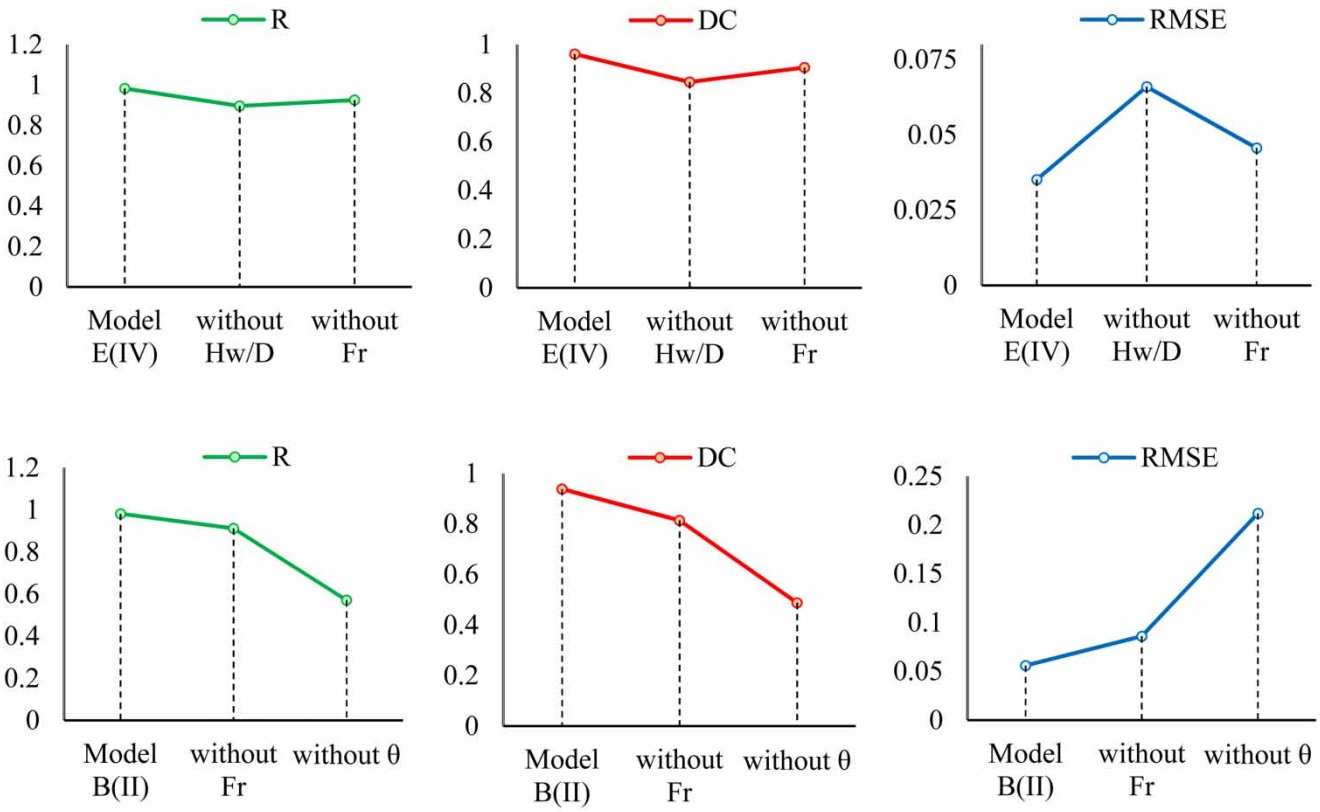

Figure 5 | Relative significance of each of input parameters of the best models.

Uncertainty is a result-dependent factor that demonstrates the range of values a modeling result can attain. It also represents the possibility that the measured value may fall into the specified range. In this study, the Monte Carlo uncertainty analysis method was used. In this procedure, two factors are considered to examine the robustness and analyze the uncertainty of the models. The first one is the percentage of the studied outputs that are in the range of 95PPU and the second one is the average distance between the upper (X_U) and lower (X_L) uncertainty bands (Noori *et al.* 2015). To do that, the considered model is developed many times (1,000 in the present study), and the empirical cumulative distribution probability of the models is obtained. After that, the X_L and X_U are considered 2.5 and 97.5% probabilities of the cumulative distribution, respectively. The appropriate confidence level is the level in which two requirements are met: (1) the 95PPU band brackets most of the observations; and (2) the average distance between the upper (at the 97.5% level) and the lower (at the 2.5% level) parts of the 95PPU is small. Quantifications of the two requirements are problem dependent to an extent. For the second requirement it is essential that the average distance between the upper and the lower 95PPU be smaller than the standard deviation of the measured data (Abbaspour *et al.* 2007). The above two indices were used to quantify the strength of calibration, accounting for the combined parameter, model, and input uncertainties. To evaluate the average width of the confidence interval band, the band width indicator was suggested by Abbaspour *et al.* (2007) as follows:

$$\text{d} - \text{factore} = \frac{\overline{dx}}{\sigma x} \tag{10}$$

where σx is the standard deviation of observed data and $\overline{dx}$ is the confidence band's average width. The percentage of the data within the confidence band of 95% is determined as follows:

$$\text{Bracketed by 95PPU} = \frac{1}{k} Cont(j|X_L < X_{reg} < X_U \tag{11}$$

where 95PPU indicates 95% predicted uncertainty; k is the number of observed data; l is the current month which changes from 1 to k and X_{reg} is the current registered data.

The results are shown in Figure 6. According to Figure 6, and based on the d-factor and 95PPU values, it can be seen that the *Fr* parameter has higher uncertainty compared with the Hw/D and θ parameters. The *Fr* parameter higher uncertainty is due to the high value of d-factor (i.e. 0.48 in bend loss state and 0.72 in inlet loss stat) and lesser value of 95PPU (i.e. 58.33 in bend loss state and 68.78 in inlet loss state).

For investigating the main effects of parameters quantitatively, the factorial analysis (FA) was also performed. FA is originated from experimental design to explore both the main and interaction effects of several factors on a response variable (Tezcan *et al.* 2015). It is particularly useful when there is a curvilinear relationship between design factors and the response variable. In fact, FA attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables. It is often used in data reduction to identify a small number of factors that explain most of the variance that is observed in a much larger number of manifest variables. FA can also be used to generate hypotheses regarding causal mechanisms or to screen variables for subsequent analysis. The results of FA are listed in Table 6. According to the results, it could be seen that the correlation coefficients between K and θ (in bend loss sate), and between K and Hw/D (in local loss state) are higher than other parameters. Therefore, the θ and Hw/D variables are more effective in energy losses modeling.

Combined data

For evaluating the performance of the GPR method for a wide range of data, data series of entrance loss were combined. Two states were considered in the data combining process: pairwise mixing and mixing all data series. Then, for predicting *Ke*, the superior model of Scenario 2 (the model E(IV)) was re-run for the mixed data. The results of this state are given in Figure 7. It could be seen that pairwise mixing of thin-wall projecting and mitered flush to 1.5:1 (horizontal to vertical) fill slope data sets led to better results in comparison with another pairwise mixing. However, according to Figure 7, the results revealed that using the mixed data set decreased the model accuracy, especially for the state of combining all

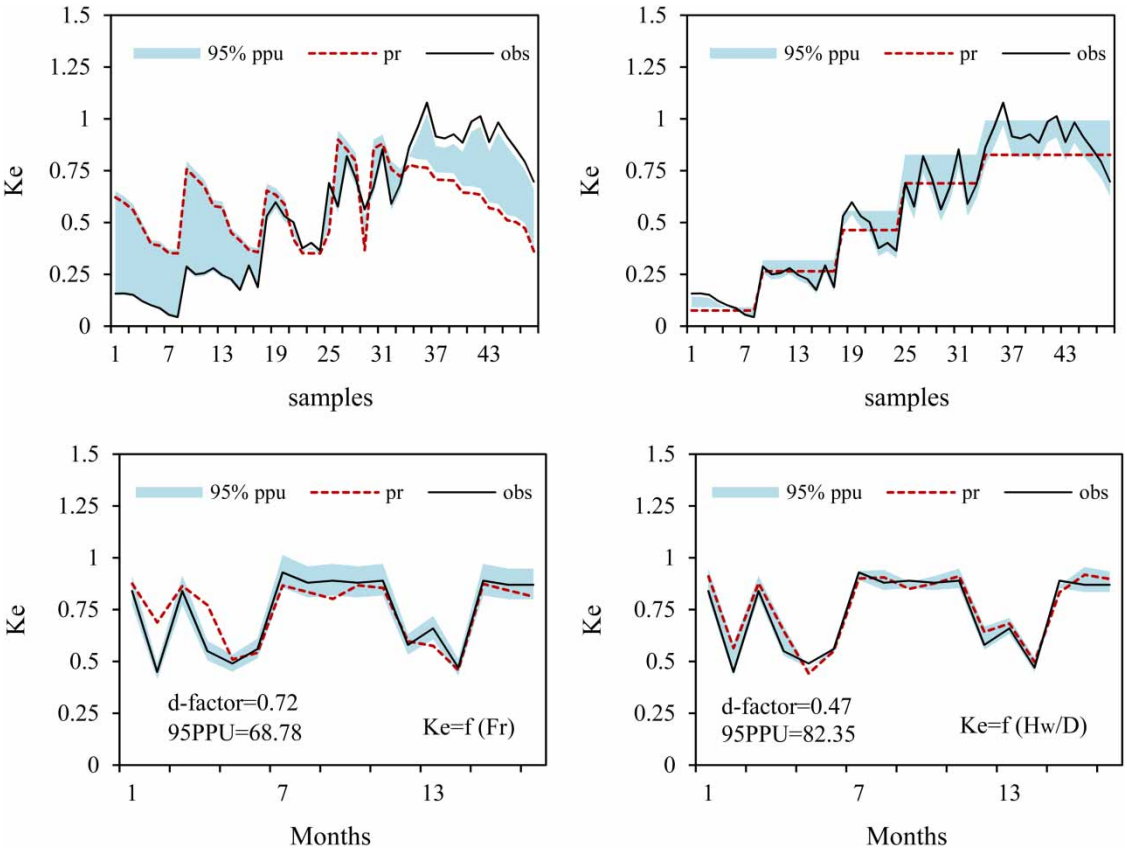

Figure 6 | Uncertainty analysis for best model of (a): scenario 1 (bend loss), and (b): scenario 2 (inlet loss).

Table 6 | The correlation coefficients between the samples of the parameters

State	Parameters			
Best model of bend loss in rectangular culvert		Fr	θ	K
	Fr	1	–0.301	–0.485
	θ		1	0.947
	K			1
Best model of inlet loss in circular culverts		Fr	Hw/D	K
	Fr	1	–0.719	–0.779
	Hw/D		1	0.947
	K			1

data series, it could be seen that the error criteria increased significantly. For all mixed data sets the values of R and DC decreased and the RMSE values increased. However, it should be noted that the models based on mixed data sets are able to cover a wider range of data and in this case, entrance loss can be studied without regarding the end inlet treatment shape.

CONCLUSIONS

In the current study, the capability of the GPR, SVM, and GEP approaches were assessed for predicting local loss in culverts. The culvert experimental data with a different shape was applied for training and testing the models. In the model development process, two scenarios were considered and bend and inlet end treatment losses were evaluated. According to the results, it was found that in Scenario 1, which investigated the bend loss in a rectangular culvert, the model with input parameters of Fr and θ led to more accurate results. It was observed that using the Froude number of the bend downstream caused an increment in model efficiency. Also, the bend upstream Froude number did not lead to undesirable results, therefore, this parameter can be used when there is no information about flow conditions of bend downstream. It showed that the bend angle had a significant impact on local loss prediction

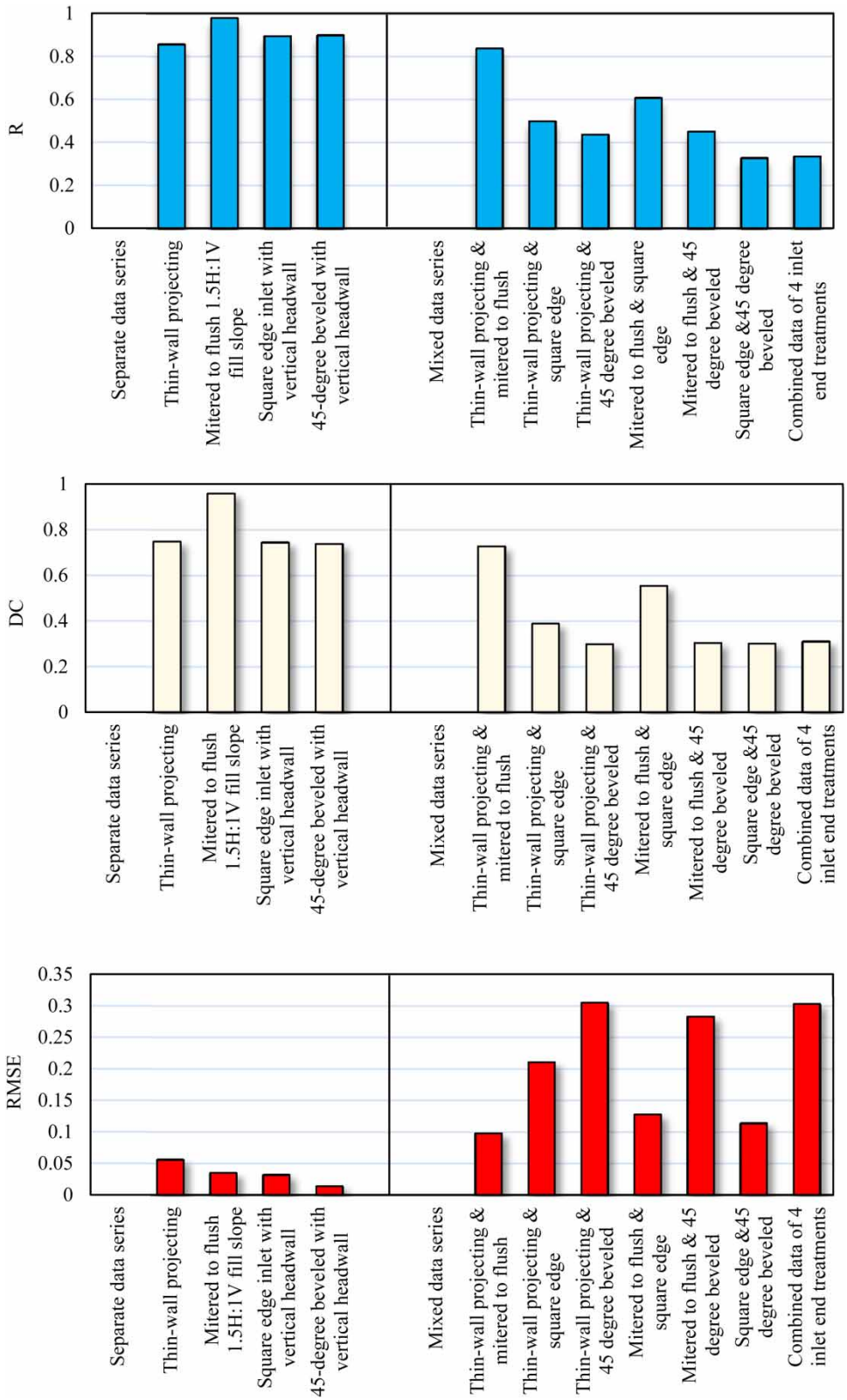

Figure 7 | The statistical parameters of the E(IV) model for separate and mixed data series.

process. The superior performance for Scenario 2 and for all end inlet treatments was obtained for the model E(IV), in which the inputs were *Fr* and *Hw/D*. It was observed that for modeling entrance loss in culverts, using relative flow depth as the input parameter improved the efficiency of the models. For models with only one input variable, using the *Fr* number led to better prediction than the *Re* number. Among the four end inlet treatments, culverts with a mitered flush to 1.5:1 (horizontal to vertical) fill slope yielded more accurate prediction. It was also observed that the mixed data set led to a less accurate outcome. From the obtained results of OAT and factorial sensitivity analysis and Monte Carlo uncertainty analysis, it was found that the correlation coefficients between K and θ (in bend loss state), and between K and Hw/D (in local loss state) were higher than other parameters. Therefore, the variable θ and Hw/D had the most significant impact on local loss prediction. Also, the *Fr* parameter had higher uncertainty compared with Hw/D and θ parameters. The proposed approaches were found to be able to predict local loss in different shaped culverts successfully, however, it should be noted that the used methods are data-driven models and the SVM, GPR, and GEP-based models are data sensitive, so further studies using data ranges out of this study and field data should be carried out to find the merits of the models to estimate local energy loss in real conditions of flow.

REFERENCES

Abbaspour, K. C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J. & Srinivasan, R. 2007 Modelling hydrology and water quality in the prealpine/alpine Thur watershed using SWAT. *J. Hydrol.* **333** (2), 413–430.

Amaranto, A., Munoz-Arriola, F., Corzo, G., Solomatine, D. P. & Meyer, G. 2018 Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland. *J. Hydroinform.* **20** (6), 1227–1246.

Anderson, D. S. 2006 *Inlet Loss Coefficients and Inlet Control Head-Discharge Relationships for Buried-Invert Culverts and Slip-Lined Culverts*. PhD thesis, Utah State University, Utah.

Azamathulla, H. M. & Wu, F. C. 2011 Support vector machine approach for longitudinal dispersion coefficients in natural streams. *Appl. Soft Comput.* **11**, 2902–2905.

Azamathulla, H. M., Haghiabi, A. H. & Parsaie, A. 2017 Prediction of side weir discharge coefficient by support vector machine technique. *Water Sci. Technol. Water Supply* **16** (4), 1002–1016.

Babovic, V. 2009 Introducing knowledge into learning based on genetic programming. *J. Hydroinform.* **11** (3–4), 181–193.

Carvalho, J., Santos, J. P. V., Torres, R. T., Santarém, F. & Fonseca, C. 2018 Tree-based methods: concepts, uses and limitations under the framework of resource selection models. *J. Environ. Inform.* **32** (2), 112–124.

Ferreria, C. 2001 Gene expression programming: a new adaptive algorithm for solving problems. *J. Complex Syst.* **13** (2), 87–129.

Ghazanfari-Hashemi, S., Etemad-Shahidi, A., Kazeminezhad, M. H. & Mansoori, A. R. 2011 Prediction of pile group scour in waves using support vector machines and ANN. *J. Hydroinform.* **13** (4), 609–620.

Gill, M. K., Asefa, T., Kemblowski, M. W. & McKee, M. 2006 Soil moisture prediction using support vector machines 1. *JAWRA J. Am. Water Resour. Assoc.* **42** (4), 1033–1046.

Holvoet, K., van Griensven, A., Seuntjens, P. & Vanrolleghem, P. A. 2005 Sensitivity analysis for hydrology and pesticide supply towards the river in SWAT. *Phys. Chem. Earth A/B/C* **30** (8–10), 518–526.

Huang, F., Huang, J., Jiang, S. H. & Zhou, C. 2017 Prediction of groundwater levels using evidence of chaos and support vector machine. *J. Hydroinform.* **19** (4), 586–606.

Kotowski, A., Szewczyk, H. & Ciezak, W. 2011 Entrance loss coefficients in pipe hydraulic systems. *Environ. Protect. Eng.* **37** (4), 105–117.

Kuss, M. 2006 *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. PhD thesis, Technischen Universität, Darmstadt.

Malone, T. R. & Parr, A. D. 2008 *Bend LOsses in Rectangular Culverts (No. K-TRAN: KU-05-5)*. Kansas Department of Transportation, Kansas.

Meshgi, A., Schmitter, P., Chui, T. F. M. & Babovic, V. 2015 Development of a modular streamflow model to quantify runoff contributions from different land uses in tropical urban environments using genetic programming. *J. Hydrol.* **525**, 711–723.

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

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? *Hydraul. Eng.* **142** (1), 04015039.

Pal, M., Singh, N. K. & Tiwari, N. K. 2014 Kernel methods for pier scour modeling using field data. *J. Hydroinform.* **16** (4), 784–796.

Rasmussen, C. E. & Williams, C. K. I. 2006 *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Roushangar, K. & Ghasempour, R. 2018 Evaluation of the impact of channel geometry and rough elements arrangement in hydraulic jump energy dissipation via SVM. *J. Hydroinform.* **21** (1), 92–103.

Smola, A. J. 1996 *Regression Estimation with Support Vector Learning Machines*. Master's thesis, Technische Universität München, Germany.

Tayyebi, A., Tayyebi, A. H., Pekin, B. K., Omrani, H. & Pijanowski, B. C. 2017 Modeling historical land use changes at a regional scale: applying quantity and locational error metrics to assess performance of an artificial neural network based back-cast model. *J. Environ. Inform.* **31** (2), 74–86.

Tezcan, U. U., Ates, F., Erginel, N., Ozcan, O. & Oduncu, E. 2015 Adsorption of disperse range 30 dye onto activated carbon derived from Holm Oak (Quercus Ilex) acorns: a 3 k factorial design and analysis. *J. Environ. Manage.* **155**, 89–96.

Tullis, B. P. 2012 *Hydraulic Loss Coefficients for Culverts*, Vol. 734. Transportation Research Board, NCHRP Report 734, Washington, DC.

Tullis, B., Robinson, S. & Young, J. 2005 Part 3: hydrology, hydraulics, and water quality: hydraulic characteristics of buried-invert elliptical culverts. Transportation research record. *J. Transport. Res. Board* **1904**, 104–112.

Vapnik, V. 1995 *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, pp. 1–47.

Yu, X., Liong, S. Y. & Babovic, V. 2004 EC-SVM approach for realtime hydrologic forecasting. *J. Hydroinform.* **6** (3), 209–223.

Zhu, S., Luo, X., Xu, Z. & Ye, L. 2018 Seasonal streamflow forecasts using mixture-kernel GPR and advanced methods of input variable selection. *Hydrol. Res.* **50** (1), 200–214.

First received 15 January 2019; accepted in revised form 22 August 2019. Available online 11 October 2019