



Exploring the application of machine learning techniques for prediction of infiltration rate

Siraj Muhammed Pandhiani¹

Received: 21 October 2021 / Accepted: 13 April 2022 / Published online: 27 May 2022
© Saudi Society for Geosciences 2022

Abstract

Water serves as an essential source for producing hydrological energy and sustainable irrigation systems, and therefore, it should be managed effectively. Attempts to manage it through conventional approaches are gradually becoming less effective due to the growing population and globally changing weather conditions. Therefore, this research investigates the suitability of support vector machine (SVM), M5P, Gaussian process (GP), random forest (RF), and multiple linear regression (MLR) methods for modeling of infiltration rate of the soil as effective and efficient substitute methods and their performances were compared with the empirical model: Kostiakov model. The performance of these models was analyzed by looking into performance measuring methods including the Nash–Sutcliffe efficiency coefficient (E), coefficient of correlation (R), and the root-mean-squared error (RMSE). The field dataset for this model contained 126 observations, 70% for the purpose of training and the remaining 30% for testing. In principle, it can be derived that the RF-based models perform better than in comparison to Nash–Sutcliffe model efficiency equal to 0.9783 and 0.9223 for the training and testing stages, respectively. Additionally, another worthy observation is that the GP, SVM, MLR, and M5P models give a good prediction performance as well, in comparison to the Kostiakov model which is inferior. According to sensitivity analysis, we are able to conclude that the most important parameter in forecasting the soil infiltration rate is time (t).

Keywords Minidisc infiltrometer · Support vector regression · Gaussian process · M5P · Random forest

Introduction

The infiltration is the process by which water is entered into the earth's surface (Hillel and Baker 1988, Barua et al. 2021). It is one of the hydrological cycle's dynamic processes. The texture of the soil, moisture content, field density, humidity, precipitation intensity, and impurity variety all have an impact on the soil's infiltration rate. The amount of infiltrated water into the soil as a result of rainfall or irrigation is important in water resource management (Sampson et al. 2021, Bouatia et al. 2020, Angelaki et al. 2013, Ostad et al. 2017). For accurate surface runoff forecasts,

the knowledge of infiltration rates is very important (Sihag et al. 2017a). The hydraulic qualities of the soil must be considered while designing drainage systems (Brooks and Corey 1964). Infiltration characteristic is one of the most significant considerations in determining the flooding situation at the catchment level (Bhave and Sreeja 2013). The ability of the soil to hold water varies depending on the texture and physical properties of the soil. Sandy soil has a faster infiltration rate and a lower water land capacity than clayey soil due to its larger pore size (Smith 2006). Additionally, surface runoff estimation is also important for hydraulic structure design and water resource development (Al-Ghobari et al. 2020, Islam and Hasan 2020). Groundwater regenerates and surface runoff is distinguished by the infiltration process from precipitation or irrigation water (Haghighi et al. 2011). The importance of the infiltration process in water resource engineering has been a good reason to encourage researchers and water resource scientists to construct several models (e.g.,

Responsible Editor: Broder J. Merkel

✉ Siraj Muhammed Pandhiani
pandhianis@ucj.edu.sa; pandhiani@hotmail.com

¹ Department of General Studies, Jubail University College, Jubail 10074, Kingdom of Saudi Arabia

Sepahvand et al. 2021; Singh et al. 2019, Ostad et al. 2020, Holtan 1961, Singh and Yu 1990, Kostiaikov 1932, Horton 1941). The three types of infiltration models are empirical models, physical models, and semi-empirical models. Mishra et al. (2003) has observed 14 unique infiltration models. Information was collected in the lab, and field studies were performed in India and the United States. The semi-empirical models (Singh-Yu, Holtan, and Horton) outperformed the other models. All three of these models have site-specific performance, but none of the models so far have been flexible, whereas Sihag et al. (2017a) compared four infiltration models (Kostiaikov, modified Kostiaikov, SCS, and novel model) for the NIT Kurukshetra campus in Haryana. The comparison showed that the data from experiential infiltration aligns better with the proposed model. Additionally, some researchers have also used studies in the area of soft computing and found success (Sihag et al. 2017a, b, c; Pandhiani et al. 2020, Pandhiani 2022, Sihag et al. 2017c; Rahmati 2017; Siddiqi et al. 2021).

Soft computing methods such as SVM, Gaussian process, M5 model tree, and random forest have been incorporated in civil and hydraulics applications in recent years and found to be very successful (Singh et al. 2017, 2021; Tiwari et al 2017). In the search of finding out the potential of soft computing techniques, Singh et al. (2021) examined that random forest is the most accurate technique in predicting the infiltration rate of the soil. Additionally, Singh et al. (2021) also examine a technique for the prediction of cumulative infiltration and infiltration rate bringing light to tree-based-soft computing. The research concludes that the random forest is the technique which is the superior out of Random Tree and M5P tree. However, Singh et al. (2019) managed to use soft computing techniques on a dataset of 132 observations which led to the conclusion that the M5P model tree is the best suitable model which can be used in the prediction of infiltration rate. Even earlier, four soft computing techniques were explored to predict infiltration rate by Sihag et al. 2018, the Gaussian process (GP), gene expression programming (GEP), and generalized neural network (GRNN). As a result of the comparison that was made, out of the four soft computing techniques, ANN achieved the first rank (correlation coefficient value 0.9816 and 0.9133 for the preparing and testing data respectively). SVM is the most accurate soft computing technique in the prediction of infiltration rate according to Vand et al.'s (2018) performed experiments in Iran.

To summarize, there is a substantial quantity of research on soil infiltration rates in literature, but just a few publications examine soft computing approaches utilized in making infiltration estimations. As a result, the primary objective of this article is to examine the possibilities of artificial intelligence methods (SVM, M5P, GP, and RF)

in predicting infiltration rates. The efficiency of these artificial intelligence techniques was also compared to those of traditional methods.

Support vector machine

SVM is a type of supervised machine learning technique that may be used to address problems like classification and regression; however, it is primarily used to solve classification problems. Based on statistical machine learning theory, in recent years, the most useful tool for forecasting was introduced by Boser et al. (1992).

The SVM method's main concept is to employ a linear platform to create nonlinear class borders, which are then used to show the input image in a high-dimensional feature space via nonlinear mapping. In the latent space, the linear model developed in the new space will reflect a nonlinear decision boundary. SVM is a learning method for classification and regression that aims to reduce classification and fitness function errors in general. In contrast to empirical risk minimization, the SVM is based on a structural risk minimization training technique. In 1998, Vapnik presented a concise overview of the theory of support vector regression.

The purpose of SVR is to evaluate a $g(P)$ function for training patterns (P) that is as far away from the training target values as possible (Q). To put it another way, SVR is a model for fitting a tube with a radius of data, resulting in the test dataset having the least amount of error. Assume that when a training set (S) is compared to its projection values, the following occurs:

$$S = \{(P_1, Q_1), (P_2, Q_2), \dots, \dots, \dots, (P_n, Q)\}, Q \in IR^n, P \in IR \quad (1)$$

where P_i is an n -dimensional vector with each element corresponding to a different decision variable. In the above question, m represents the number of samples, and Q_i is the equivalent output variable. According to Vapnik, to decrease the test error, the term displaying the complexity of a group of functions must be minimized (1). As a result, the SVR technique uses linear functions in the classical sense such as $g(P) = w \cdot N + b$ (i.e., b and w are bias value and weight vectors, respectively) for the forecasting.

The predictions differ from the measurements statistically, and in some circumstances, error levels less than γ are impossible to consider. In conclusion, the deviation from the slack variable (ψ_i^+ and ψ_i^-), known as γ , is defined. The following Eq. (2) is then used to minimize the error value and maximum margin necessitates the elimination of the weight vector's norm:

$$\text{Testingerror} \leq \text{Trainingerror} + \text{Complextityofsetofmodels}$$

$$\text{Minimize : } \frac{1}{2}(w.w) + C \sum_{i=1}^n (\psi_i^+ \text{ and } \psi_i^-) \tag{2}$$

$$\text{Constraints : } \begin{cases} w.M_i + a - N_i \leq \gamma + \psi_i^+ i = 1, 2, 3, \dots, n \\ N_i - (w.M_i + a) \leq \gamma + \psi_i^- i = 1, 2, 3, \dots, n \\ \psi_i^+ \geq 0, \psi_i^- \geq 0 i = 1, 2, 3, \dots, n \end{cases} \tag{3}$$

where C adjusts the deviation error more than γ and the radius γ defines the estimated tube range, where both, C and γ , are above zero.

Gaussian process regression

The Gaussian process regression method shows how to generalize nonlinear and compound function mapping contained in data sets using a probabilistic, multivariate supervised learning technique. The GP regression is also commonly used in other engineering and scientific fields, where it is sometimes referred to as kriging, especially in cases where there are not many parameters. Originally, the method was developed to interpolate sparse data points between two dimensions in geostatistics. Now, it is widely used in two- and three-dimensional spatial mapping in the geological and meteorological fields. Recent developments in the community of statistics and machine learning have used this method and those similar to it to solve higher order problems. In GPR models, training data is used to predict new input based on probabilistic models. In order to model the outputs, y , in terms of input parameters, x , we use Eq. (4), where $h(x)$ and β are a set of basis functions, and $f(x)$ is a Gaussian process with zero mean and covariance function $k(x, x')$.

$$y = h(x)^T \beta + f(x) \sim GP(0, k(x, x')) \tag{4}$$

Basis functions are obvious functions of parameters at a given point and are most commonly polynomials, whereas covariance functions are relationships between the required parameter point and training data away from the basis function Kuss (2006).

The basis function represents the exact relationship between the parameters at one point, which is usually a simple polynomial, while the covariance function defines the relation between the required parameter point and the training data points away from the basis function. There is zero mean in the covariance function, and it varies around the values of the parameters derived from the basis function. To determine the model parameters, the model must be fitted to the data. These are the coefficients of the basis function, the covariance function, and the noise level in the data Rasmussen and Williams (2006). The GP regression model, according to Rasmussen and Williams (2006), works

because adjoining observations can communicate evidence about each other. It is the process of defining a prior across function space in real-time. The Gaussian distribution's mean and covariance are both vectors and matrices, and the Gaussian function is just over the function. The predictive distribution that is equal to the test input is recognized by the GP regression model. The GP technique is a set of random variables with a joint multivariate Gaussian distribution for any finite number. Consider PXQ denote the input and output domains, respectively, from which n pairs (P_i, Q_i) are distributed independently and identically. For regression, let $N \subseteq \mathfrak{R}$; then, a GP on χ is defined by a mean function $\mu : \chi \rightarrow \mathfrak{R}$ and a covariance function $k : \chi \times \chi \rightarrow \mathfrak{R}$. In addition, Kuss (2006) is a good resource for additional information on GP regression with various covariance functions.

Model tree (M5P)

Model trees were first proposed by Quinlan (1992), and then, Wang (1997) rebuilt and enhanced the concept in the M5P system. An M5P model tree is an effective learning strategy for predicting real values. The large datasets benefit from model trees. The M5P model tree algorithm starts by recursively dividing the instance space to create a regression tree. The separating parameter is chosen to reduce intra-subset inconsistency in the values as they proceed from root to branch to node. The inconsistency is evaluated by determining the predicted error decrease as a result of testing each characteristic at a similar node, and then via the branch, derive the node from the root using the value of standard deviation. The feature with the largest anticipated error decline is chosen. When the total number of cases reaching a node is relatively low, or only a few other cases remain, the unbearable comes to an end. Equation (5) is used to calculate the standard deviation reduction (SDR):

$$SDR = sd(P) - \sum \frac{|P_j|}{|P|} sd(P_j) \tag{5}$$

where P is fixed of patterns that enter the node, P_j is the subsection of patterns with the j th prospective set conclusion, and the standard deviation is denoted by sd (Wang 1997).

Random forest

The random forests, also known as random decision forests, are a form of collective learning method for classification, a regression that uses a combination of tree estimators to create each tree, with each tree

being produced using a random vector that is independently tested from the input vector. The tree estimator applies arbitrary numerical values to the class labels of the random forest classifier in regression (Breiman 1999). Random forest regression is also utilized in this study to expand a tree by employing an arbitrarily chosen variable or a combination of variables at each node. There is a tree estimator’s plan that necessitates the selection of a variable choice metric. Various techniques to deal with the variable’s choice for tree acceptance are proposed in the records, with the majority explicitly giving a quality metric to the variable. The information gain ratio criteria and the Gini index are the most commonly used variable choice measures in tree induction. (Quinlan 1992; Breiman et al. 1984).

Variables at each node (n) need to be recognized to construct a tree along with the total number of trees that are to be built (k); Breiman (1999) provided these two user-defined parameters for random forest regression. Only a certain number of variables can be established at each node in order to obtain the best split. In conclusion, it is up to the consumer to define the number of trees (k) for which the random forest regression has to be developed and the output values of the random forest-based regression are numeric.

Empirical models

Using the least-square technique on training data sets, regression coefficients were derived for the Kostiakov and multilinear regression models that are known as empirical models.

Multiple linear regression

MLR is a well-known technique for predicting the values of each independent variable. MLR also calculates the relevance of a variable by looking at the connection between infiltration and environmental variables. Multiple predictor parameters are evaluated to MLR. The structure of the conventional MLR model is:

$$M = c_0 a_1^{c_1} a_2^{c_2} a_3^{c_3} a_4^{c_4} \dots \dots \dots a_n^{c_n} \tag{6}$$

where M is the standard value expressed as a function of the number of independent parameters, and n is the number of independent parameters $a_1, a_2, a_3, \dots, a_n$, in which the values of coefficients, $c_0, c_1, c_2, c_3, \dots, c_n$, are unknown. The least-square method (LSM) is used to evaluate these values, which correspond to the actual performance.

Table 1 Properties of the material selected for experimentation

Properties	S	Fa	C
Specific gravity	2.48	2.07	1.59
D_{50}	0.438	0.180	
C_u	3.1290	2.7333	
Color	White	Gray	Brownish yellow

Table 2 Information of the soil samples with their moisture content

S (%)	C (%)	Fa (%)	Mc (%)
45	45	10	2,5,10,15,20
40	40	20	2,5,10,15,20
35	35	30	2,5,10,15,20
30	30	40	2,5,10,15,20
25	25	50	2,5,10,15,20

Kostiakov model

The following observed model was proposed by Kostiakov (1932) for estimating the soil infiltration rate:

$$g(t) = pt^{-q} \tag{7}$$

In the above mentioned Eq. (7), where t is denoted by the time of infiltration (T), $g(t)$ is the rate of infiltration at time $t(LT - 1)$, and p and q are dimensionless observed constants.

Methodology and data set

All of the experiments were carried out in a hydraulics laboratory at the National Institute of Technology in Kurukshetra, India, using a mini-disk infiltrometer. The mini-disk infiltrometer consists of two chambers. The first is a reservoir of water, and the second is a bubble. Through a Mariotte tube, both are connected. An ideal water pressure head of 0.05 to 0.7 kPa can be achieved with this tube. A sintered steel disk is incorporated into the bottom part of the instrument, which has a diameter of 4.5 cm and a thickness of 3 mm. Both chambers are filled with water, so water moves into the soil from the flat surface. The reservoir chamber was measured at specific intervals by recording the amount of water in it. The soil used to calculate the infiltration rate contains varying amounts of sand (S), clay (C), and fly ash (F) (Fa). A proctor of volume 1000 cm³ was selected for the compaction of samples. All of the initial requirements, such as moisture content (Mc) and dry density, were planned ahead of time. The parameters of the soil samples are shown in Tables 1 and 2, as well as data on how soil samples are mixed and their moisture contents.

Dataset

In this study, the experiments were held in the NIT Kurukshetra laboratory, where 126 datasets were produced. According to our definitions, 88 datasets were randomly selected from a total of 126 for model training, while the leftover 38 datasets were selected for model testing. The input dataset includes time (*t*) in seconds, clay (*C*) in percent, sand (*S*) in percent, fly ash (*Fa*) in percent, bulk density (*Bd*) in gm/cc, and moisture content (*Mc*) in percent, while the output dataset includes infiltration rate (*g(t)*) in mm/h where Table 3 details the characteristics of the experimental data, where descriptive statistics were for testing the data validity. The process of model construction entails splitting data into two sections expending the trial-and-error technique. The complete data was split into two categories: training (70%) and testing (30%). The first 70% of data from the entire data set were used to train the network to obtain the parameters model, and the remaining 30% was used for testing purposes to verify the descriptive statistics. Table 3 shows descriptive statistics for the infiltration rate and six subsets, such as mean (μ), standard deviation (S_x), the coefficient of variation (C_{vx}), skewness coefficient (C_{sx}), and range (maximum and minimum values). It seems from the table observed infiltration rate shows a highly skewed distribution ($C_{sx} = 2.12$ m²/h). In the training infiltration rate data, the mean, range information (5.77 m²/h-0.541 m²/h-30.659 m²/h) is higher than the testing stage data (4.82 m²/h-0.573 m²/h-24.604 m²/h), respectively. Table 3 also shows that the training dataset has more uncertainty and variability than the testing

dataset, based on the result of standard deviation (S_x), variance, and skewness coefficient. The study’s flow-chart is shown in Fig. 1.

The accuracy of the proposed model

The strength of the proposed models was measured using three well-known performance measuring methods, the coefficient of correlation (*R*), root-mean-square error (RMSE), and Nash–Sutcliffe model efficiency. Coefficient (*E*) values were calculated using training on the testing dataset. In both modeling and forecasting, the criteria for judging the best model are based on minimum RMSE values and the maximum *R* and *E* values.

$$R = \frac{\sum (p - \bar{p})(q - \bar{q})}{\sqrt{[\sum (p - \bar{p})^2][\sum (q - \bar{q})^2]}} \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - q_i)^2} \tag{9}$$

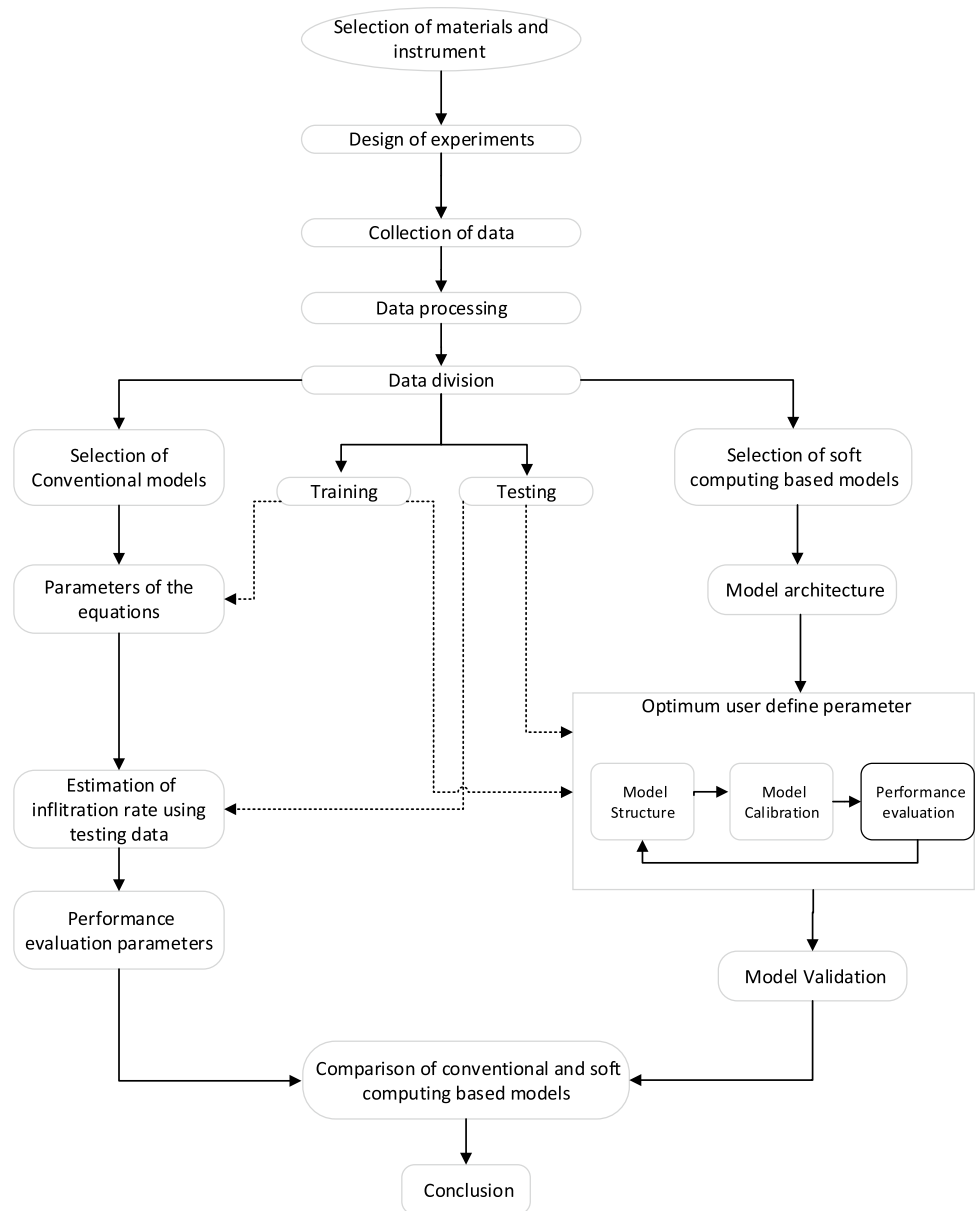
$$E = 1 - \frac{\sum_{i=1}^n (p_i - q_i)^2}{\sum_{i=1}^n (p_i - \bar{q})^2} \tag{10}$$

where in the above equation, *n* is nuber of observation, *p* is the value of the independent variable (observed value), and *q* is the value of a dependent variable (predicted value).

Table 3 Pertinent information for infiltration rate and six subsets data

Variable	Units	Range	μ	S_x	C_{vx}	C_{sx}
Training data set						
<i>t</i>	sec	87.62–9916.11	2163.18	2128.46	0.98	1.49
<i>C</i>	%	25–45	34.15	7.24	0.21	0.14
<i>S</i>	%	25–45	34.15	7.24	0.21	0.14
<i>Fa</i>	%	10–50	31.705	14.48	0.46	–0.14
<i>Bd</i>	gm/cc	1.37–1.91	1.63	0.14	0.09	–0.09
<i>Mc</i>	%	2–20	9.99	6.17	0.62	0.23
<i>f(t)</i>	mm/h	0.541–30.659	5.77	6.19	1.07	2.12
Testing data set						
<i>t</i>	s	92–8568.19	2282.03	2036.12	0.89	1.46
<i>C</i>	%	25–45	33.46	7.27	0.22	0.30
<i>S</i>	%	25–45	33.46	7.27	0.22	0.30
<i>Fa</i>	%	10–50	33.08	14.54	0.44	–0.30
<i>Bd</i>	gm/cc	1.37–1.90	1.628	0.147	0.09	–0.11
<i>Mc</i>	%	2–20	10.54	6.84	0.65	0.17
<i>f(t)</i>	mm/h	0.573–24.604	4.82	4.77	0.99	2.31

Fig. 1 Flow chart of the research



Implementation of machine learning methods

As for performance evaluation parameters, three standard statistical measures were selected: *R*, RMSE, and *E*. These measures judged the accuracy of the machine learning models and the Kostikov model. The primary parameters were tested manually several times to determine the optimal value. The model’s better prediction accuracy is indicated by higher values of *R*, *E*, and lower values of RMSE. Random forest regression involves growing the trees (*k*) in the forest and selecting the features or variables (*m*) to be applied at each node to create each tree. Model calibration in M5P

is done by changing the value of the number of instances allowed at each node (*m*). There are multiple kernel functions in GP and SVM, deciding which one is best is also a research subject on its own. Nevertheless, in this research, the radial basis kernel (RBF) and the Pearson VII function kernel (PUK) are used.

1. $RBF = (e^{-\gamma \|p_i - q_j\|^2})$
2. $PUK = \left(1 / \left[1 + \left(2\sqrt{\|p_i - q_j\|^2} \sqrt{2^{(1/\omega)} - 1} / \sigma \right)^2 \right]^\omega \right)$

The above kernel function specified well-known parameters like γ , σ and ω , where C is a parameter on regularization and the error-insensitive zone’s size ϵ are needed for SVM, while the amount of noise is required for GP regression. Both GP regression and SVR used the same kernel-specific parameters. Table 4 shows the best user-defined parameter values for different approaches.

Result and discussion

The Kostiakov model and MLR are both experiential models. Using the least-square methodology, the training dataset was used to drive regression equation coefficients in the Kostiakov and MLR models.

Kostiakov model:

$$f(t) = 900.23t^{-0.7958} \tag{11}$$

MLR:

$$f(t) = 4.112t^{-0.615}C^{0.74}S^{-0.064}F_a^{0.962}B_d^{-4.75}M_c^{0.39} \tag{12}$$

A comparative analysis of experimental results with the results obtained from the Kostiakov model is described in Fig. 2. The Kostiakov model produces poor forecasts, as shown in Fig. 2. In contrast to the measured and predicted infiltration rate values obtained from the model MLR for training and testing, samples are depicted in Fig. 3. Predicted infiltration rate values obtained from the model MLR is in closer agreement with the experimentally observed infiltration rate values.

Comparisons between the actual infiltration rate and the predictions obtained from different artificial intelligence-based models for training and testing samples are depicted in Figs. 4, 5, 6 and 7. According to these figures, the predicted infiltration rate values obtained from artificial intelligence-based models are in good agreement with the, experimentally obtained, actual infiltration rate values.

Table 5 shows that the RF model outperforms other models during the training and testing period. SVM and GP-based models work well for the training period but in the

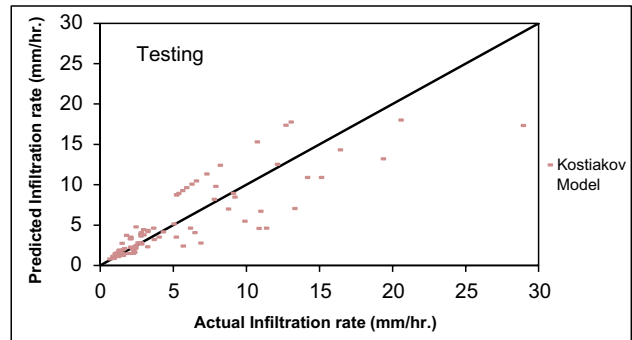
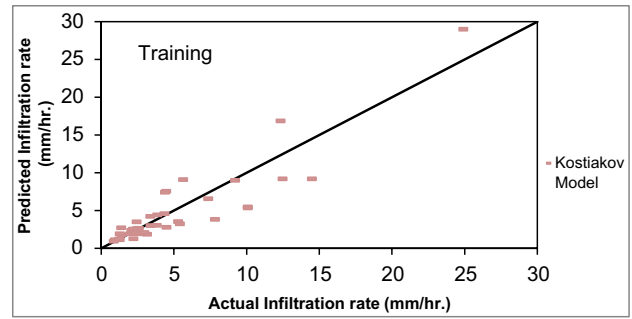


Fig. 2 Comparing the actual and predicted infiltration rates from the Kostiakov model for training and testing data

testing period, their performance is poor. Figure 7 indicates that the M5P model predicts some negative values in the training and testing period. Table 5 suggests that MLR also works better than SVM, GP, M5P, and Kostiakov models.

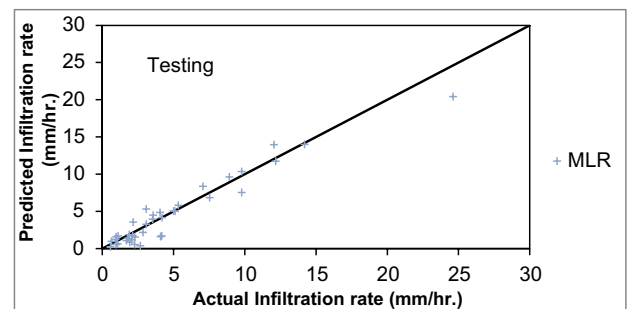
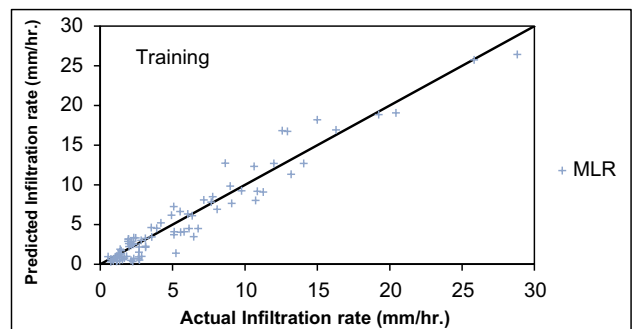


Fig. 3 Comparing the actual and predicted infiltration rates from the MLR model for training and testing data

Table 4 User-defined parameters SVM, GP, M5P, and random forest regression

Approaches	RBF kernel	PUK kernel
SVM	$C = 10, \gamma = 2$	$C = 10, \omega = 0.1, \sigma = 6$
GP	noise = 0.01, $\gamma = 2$	noise = 0.01, $\omega = 0.1, \sigma = 6$
M5P	$m = 2$	
RF	$m = 2, k = 10, I = 100$	

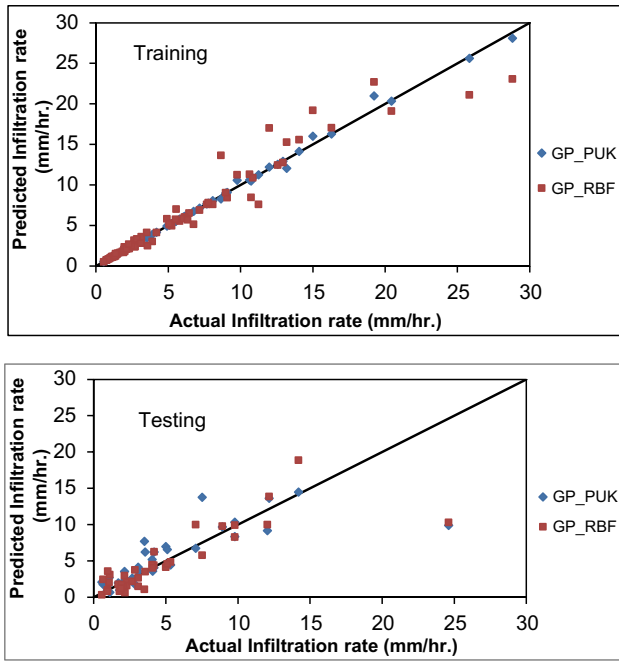


Fig. 4 Comparing the actual and predicted infiltration rates from the GP model for training and testing data

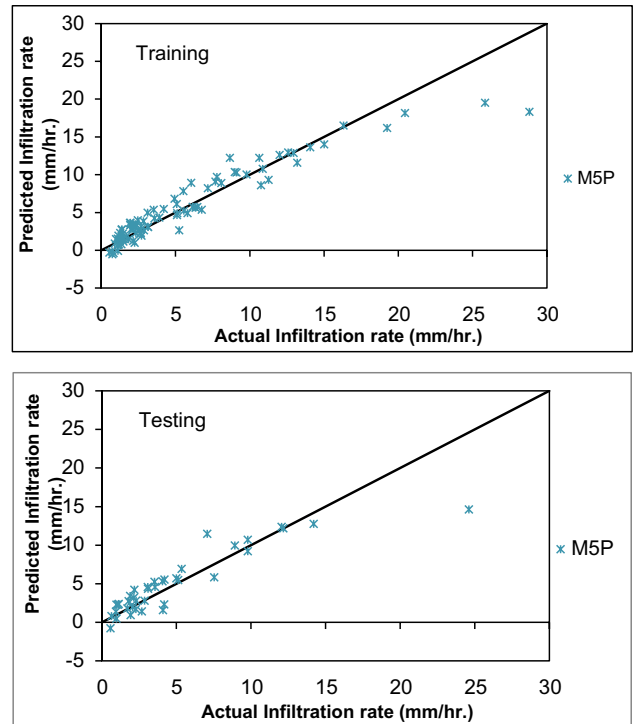


Fig. 6 Comparing the actual and predicted infiltration rates from the MP5 model for training and testing data

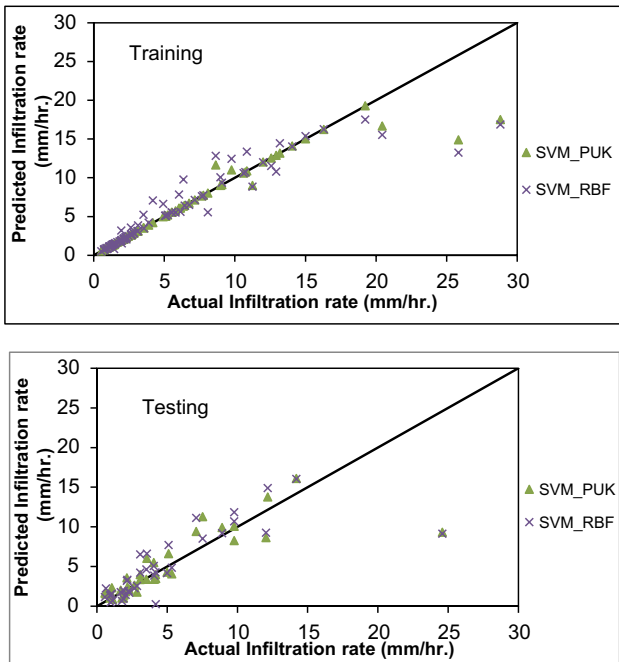


Fig. 5 Comparing the actual and predicted infiltration rates from the SVM model for training and testing data

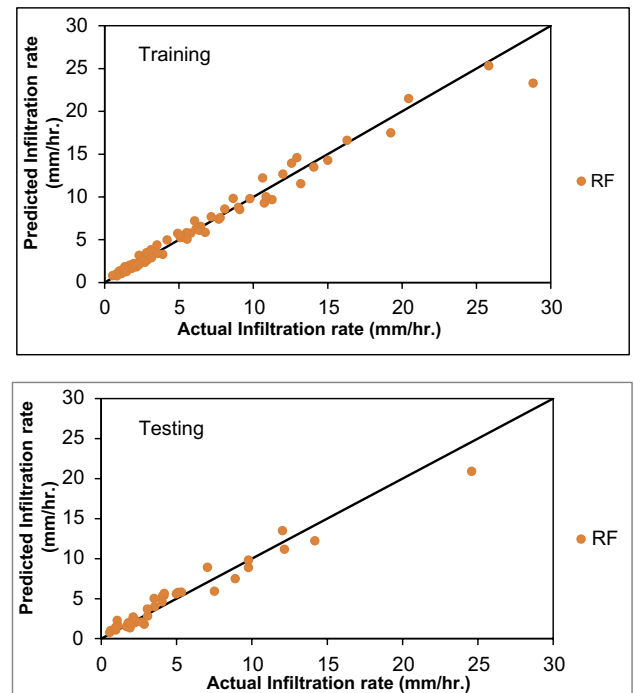


Fig. 7 Comparing the actual and predicted infiltration rates from the RF model for training and testing data

For the comparison of different artificial intelligence technique-based models, MLR and Kostiakov model's graph is plotted in Fig. 8. The line of the perfect agreement for the random forest regression model predicted values were

Table 5 Performance indicators of the GP, SVM, M5P, RF, MLR, and Kostiakov models

Approaches	Training data set			Testing data set			Ranking
	<i>R</i>	RMSE	<i>E</i>	<i>R</i>	RMSE	<i>E</i>	
GP_PUK	0.9988	0.3043	0.9976	0.7939	4.0522	0.4263	7
GP_RBF	0.9696	1.5078	0.9400	0.8212	2.7327	0.4916	5
SVM_PUK	0.9589	2.0214	0.8922	0.8095	2.7982	0.4669	6
SVM_RBF	0.9250	2.5178	0.8328	0.7839	2.9829	0.3942	8
M5P	0.9459	2.1743	0.8753	0.9006	2.1068	0.6978	3
RF	0.9907	0.9070	0.9783	0.9788	1.0681	0.9223	1
MLR	0.9715	1.5006	0.9406	0.9638	1.2918	0.8864	2
Kostiakov model	0.8796	2.9292	0.7736	0.9025	2.2058	0.6688	4

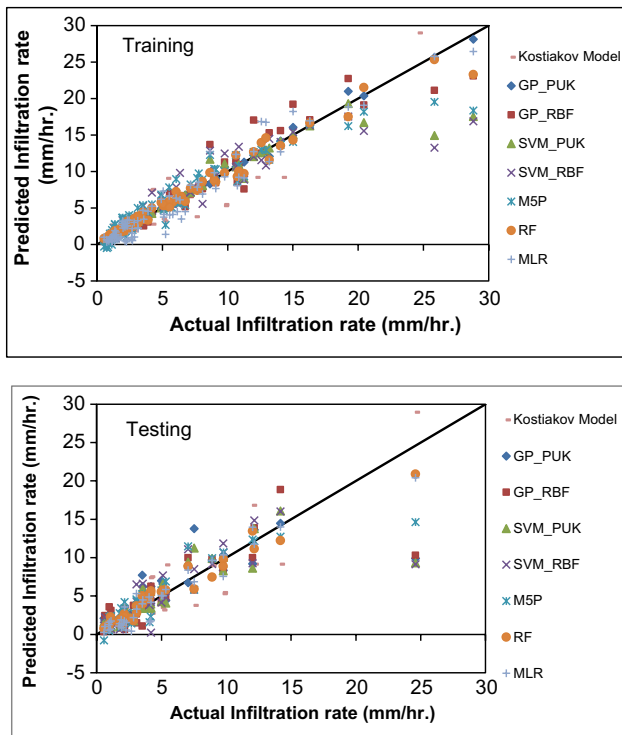


Fig. 8 Comparison of the actual and predicted infiltration rate values using various artificial Intelligence techniques based models and Kostiakov model for training and testing data

very close, as can be seen in Fig. 8. Figure 9 shows a graph between the number of test data sets and relative error, which reflects the random forest regression model’s improved performance. This graph shows that the random forest-based regression model has the least error between observed and expected values.

Figure 10 allows us to compare techniques, on basis of the correlation coefficient, mentioned earlier in the following studies: Singh et al. (2017), Sihag et al. (2018), Vand et al. (2018), and Singh et al. (2021), with the current study. As can be observed, the value of *R* for Singh et al.

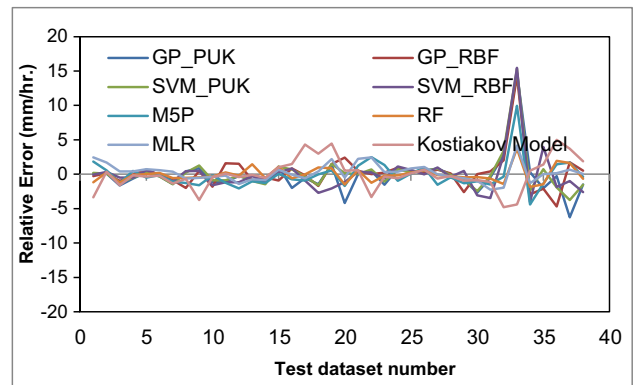


Fig. 9 Relative error between the actual and predicted infiltration rates based on different artificial Intelligence techniques based models and Kostiakov model for testing data

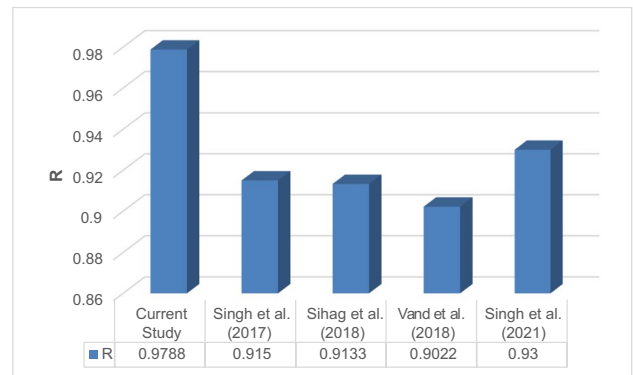


Fig. 10 Comparison with the previous studies

(2017) is 0.915, Sihag et al. (2018) is 0.9133, Vand et al. (2018) is 0.9022, and Singh et al. (2021) is 0.9300, which are all inferior when compared to the value that is of the current study (0.9788). Thus, it can be concluded that the comparison validates the model of the current study as the best one amongst those which have been put forward in previous studies as successful soft computing techniques.

Table 6 Results of sensitivity analysis of RF and MLR models

Inputs	Removed	RF Model		MLR Model	
		R	RMSE	R	RMSE
<i>t, C, S, Fa, Bd, Mc</i>	-	0.9788	1.0681	0.9006	2.1068
<i>C, S, Fa, Bd, Mc</i>	<i>T</i>	0.4283	5.0133	0.4319	4.8089
<i>t, S, Fa, Bd, Mc</i>	<i>C</i>	0.9786	1.0715	0.9006	2.1068
<i>t, C, Fa, Bd, Mc</i>	<i>S</i>	0.9786	1.0715	0.9006	2.1068
<i>t, C, S, Bd, Mc</i>	<i>Fa</i>	0.9781	1.0792	0.9006	2.1068
<i>t, C, S, Fa, Mc</i>	<i>Bd</i>	0.9784	1.0416	0.8990	2.1116
<i>t, C, S, Fa, Bd</i>	<i>Mc</i>	0.9753	1.1188	0.9005	2.0983

Sensitivity analysis

Sensitivity analysis allows the examination of how influential each of the individual parameters is to the model's ability to predict the soil's infiltration rate. The analysis of the models began after the model structure was adjusted to define the most useful parameters. In the absence of each input parameter, indicators such as *R* and RMSE were employed to assess the models' performance. The performance changed when one of the input parameters was removed and the effect of each parameter was then further investigated based on the degree of variation it caused in performance. In Table 6, the results of the sensitivity analysis of the most accurate models, RF and MLR, have been presented. As can be observed, a lack of time (*t*) results in a substantial decrease in model accuracy; therefore, we gather that the most essential parameter for predicting the soil infiltration rate was time (*t*).

Conclusion

Observations made in this article address the inability of conventional methods to provide comprehensive results in the prediction of soil infiltration rates. In order to achieve better results, we make use of AI-based, MLR, and tree algorithm applications. The performance of five different machine learning methods for the regression problem is analyzed and compared in this paper to compare responses in soil infiltration rate modeling. The results demonstrate the applied methods are suitable techniques to predict infiltration rate values. Our five model methods are multiple linear regression (MLR), support vector machines (SVM), a Gaussian process (GP), M5P, and random forests (RF). The experimental data of infiltration rate is observed on fly ash, clay, and sand mixed samples. Furthermore, on the given data set, the experiments compared the performance of tree models (M5P and random forests) with non-tree models (Gaussian process, support vector machines, and multilinear regression). It was found that the obtained results of our machine learning methods compared to the conventional

model lead to better infiltration rate errors in estimation. The obtained results indicate that the RF model performs at best to forecast the infiltration rate of soil for the given data set, with the MLR model close behind, but both models outperforming the SVM, GP, M5P, and Kostiakov models considerably. Hence why, from the findings of this article we conclude that the presented models, RF and MLR, are the most reliable and could perform as one of the most stable models available at present as far as the experiments conducted in this article. In addition, it is also worth noting that according to the sensitivity analysis, the most important parameter in determining the soil infiltration rate is time (*t*). Lastly, to build upon this article and its findings, in the future, work can be done to investigate how the capabilities of these AI models can be enhanced so they can be used to train a general model to predict the infiltration process which can then be successfully implemented in the other areas of studies along with providing even more accurate forecasts in the area discussed in the article.

Author contribution The corresponding author Siraj Muhammed Pandhiani performed all the contributions, including writing, analyzing, investigating, and supervising the article.

Declarations

Consent to participant Not applicable.

Conflict of interest The authors declare no competing interests.

References

- Al-Ghobari H, Dewidar A, Alataway A (2020) Estimation of surface water runoff for a semi-arid area using RS and GIS-based SCS-CN method *Water* 12(7):1924. <https://doi.org/10.3390/w12071924>
- Angelaki A, Sakellariou-Makrantonaki M, Tzimopoulos C (2013) Theoretical and experimental research of cumulative infiltration. *Transp Porous Media* 100(2):247–257. <https://doi.org/10.1007/s11242-013-0214-2>
- Barua S, Mukhopadhyay BP, Bera A (2021) Integrated assessment of groundwater potential zone under agricultural dominated areas in the western part of Dakshin Dinajpur district, West Bengal, India. *Arab J Geosci* 14:1042. <https://doi.org/10.1007/s12517-021-07312-y>
- Bhave S, Sreeja P (2013) Influence of initial soil condition on infiltration characteristics determined using a disk infiltrometer. *ISH J Hydraul Eng* 19(3):291–296. <https://doi.org/10.1080/09715010.2013.808445>
- Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. ACM, pp 144–152. <https://doi.org/10.1145/130385.130401>
- Bouatia M, RafikDemagh R, Derriche Z (2020) Structural behavior of pipelines buried in expansive soils under rainfall infiltration (Part I: transverse behavior). *Civil Eng J* 6(9):1822–1838. <https://doi.org/10.28991/cej-2020-03091585>

- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Monterey
- Breiman L (1999) Random forests — random features. Technical Report 567, Statistics Department, University of California, Berkeley, <ftp://ftp.stat.berkeley.edu/pub/users/breiman>
- Brooks RH, Corey AT (1964) Hydraulic properties of porous media and their relation to drainage design. *Trans ASAE* 7(1):26–0028
- Devis D (2014) Mini disk infiltrometer user's manual, Version 9. Decagon Devices, Pullman
- Haghighi F, Saghafiyan B, Kheirkhah M (2011) Evaluation of soil hydraulic parameters in soils and land use change. In: Dar IA and Dar MA (eds) *Earth and Environmental Sciences*. IntechOpen. <https://doi.org/10.5772/19059>
- Hillel D, Baker RS (1988) A descriptive theory of fingering during infiltration into layered soils. *Soil Sci* 146(1):51–56
- Holtan HN (1961) A Concept for Infiltration Estimates in Watershed Engineering. Agricultural Research Service, United States Department of Agriculture pp 41–51
- Horton RE (1941) An approach toward a physical interpretation of infiltration-capacity. *Soil Sci Soc Am J* 5(C):399–417
- Islam MA, Hasan H (2020) Generation of IDF equation from catchment delineation using GIS. *Civil Eng J* 6(3):540–547
- Kostiakov AN (1932) On the dynamics of the coefficient of water-percolation in soils and on the necessity for studying it from a dynamic point of view for purposes of amelioration. *Trans* 6:17–21
- Kuss M (2006) Gaussian process models for robust regression, classification, and reinforcement learning (Technische Universität Darmstadt, Darmstadt, Germany). PhD Thesis published online. <https://tuprints.ulb.tu-darmstadt.de/id/eprint/674>
- Mishra SK, Tyagi JV, Singh VP (2003) Comparison of infiltration models. *Hydrol Process* 17(13):2629–2652. <https://doi.org/10.1002/hyp.1257>
- Ostad AA, Shayannejad M (2020) Impermanent changes investigation of shape factors of the volumetric balance model for water development in surface irrigation. *Modeling Earth Systems and Environment*. Springer Nature Switzerland AG., 6(3):1573–1580. <https://doi.org/10.1007/s40808-020-00771-4>
- Ostad AA et al (2017). Deficit irrigation: optimization models. Management of drought and water scarcity. *Handbook of drought and water scarcity*, Chapter No. 18, Vol. 3, pp: 373–389. Taylor & Francis Publisher. Imprint: CRC Press. eBook ISBN: 9781315226774. 1st Edition. <https://doi.org/10.1201/9781315226774>
- Pandhiani SM (2022) Assessment of coefficient of discharge of gabion weir using soft computing techniques. in press. *Int J Hydrol Sci Technol*. <https://doi.org/10.1504/IJHST.2021.10043279>
- Pandhiani SM, Sihag P, Shabri AB, Singh B, Pham QB (2020) Time-series prediction of streamflows of Malaysian rivers using data-driven techniques. *J Irrig Drain Eng., ASCE Library* 146(7):1–12. <https://ascelibrary.org/doi/10.1061/%28ASCE%29IR.1943-4774.0001463>
- Quinlan JR (1992) Learning with continuous classes. *Proceedings of Australian Joint Conference on Artificial Intelligence*. World Scientific Press, Singapore, pp 343–348
- Rahmati M (2017) Reliable and accurate point-based prediction of cumulative infiltration using soil readily available characteristics: a comparison between GMDH, ANN, and MLR. *J Hydrol* 551:81–91. <https://doi.org/10.1016/j.jhydrol.2017.05.046>
- Rasmussen CE, Williams CK (2006) *Gaussian processes for machine learning*, vol 1. MIT press, Cambridge
- Sampson AP, Weli VE, Nwagbara MO, Eludoyin OS (2021) Sensations of air temperature variability and mitigation strategies in urban environments. *J Hum Earth Futur* 2(2):100–113
- Sepahvand A, Singh B, Ghobadi M et al (2021) Estimation of infiltration rate using data-driven models. *Arab J Geosci* 14:42. <https://doi.org/10.1007/s12517-020-06245-2>
- Siddiqi TA, Ashraf S, Khan SA et al (2021) Estimation of data-driven streamflow predicting models using machine learning methods. *Arab J Geosci* 14:1058. <https://doi.org/10.1007/s12517-021-07446-z>
- Sihag P, Tiwari NK, Ranjan S (2017a) Modelling of infiltration of sandy soil using Gaussian process regression. *Model Earth Syst Environ* 3(3):1091–1100. <https://doi.org/10.1007/s40808-017-0357-1>
- Sihag P, Tiwari NK, Ranjan S (2017b) Prediction of unsaturated hydraulic conductivity using an adaptive neuro-fuzzy inference system (ANFIS). *ISH J Hydraul Eng*: 1–11. <https://doi.org/10.1080/09715010.2017.1381861>
- Sihag P, Tiwari NK, Ranjan S (2017c) Estimation and inter-comparison of infiltration models. *Water Sci* 31(1):34–43. <https://doi.org/10.1016/j.wsj.2017.03.001>
- Sihag P, Singh B, Sepah Vand A, Mehdipour V (2018) Modeling the infiltration process with soft computing techniques. *ISH J Hydraul Eng*: 1–15. <https://doi.org/10.1080/09715010.2018.1439776>
- Singh VP, Yu FX (1990) Derivation of infiltration equation using a systems approach. *J Irrig Drain Eng* 116(6):837–858. [https://doi.org/10.1061/\(ASCE\)0733-9437\(1990\)116:6\(837\)](https://doi.org/10.1061/(ASCE)0733-9437(1990)116:6(837))
- Singh B, Sihag P, Singh K (2017) Modelling of the impact of water quality on infiltration rate of soil by random forest regression. *Model Earth Syst Environ* 3(3):999–1004. <https://doi.org/10.1007/s40808-017-0347-3>
- Singh B, Sihag P, Singh K (2018) Comparison of infiltration models in NIT Kurukshetra campus. *Appl Water Sci* 8(2):63. <https://doi.org/10.1007/s13201-018-0708-8>
- Singh B, Sihag P, Deswal S (2019) Modelling of the impact of water quality on the infiltration rate of the soil. *Appl Water Sci* 9:15. <https://doi.org/10.1007/s13201-019-0892-1>
- Singh B, Sihag P, Parsaie A, Angelaki A (2021) Comparative analysis of artificial intelligence techniques for the prediction of infiltration process. *Geol Ecol Landsc* 5(2):109–118. <https://doi.org/10.1080/24749508.2020.1833641>
- Smith B (2006) *The farming handbook*. University of Kwazulu-Natal Press and CTA, Wageningen, pp 37–132
- Tiwari NK, Sihag P, Ranjan S (2017) Modeling of infiltration of soil using adaptive neuro-fuzzy inference system (ANFIS). *J Eng Technol Educ* 11(1):13–21
- Vand AS, Sihag P, Singh B, Zand M (2018) Comparative evaluation of infiltration models. *KSCE J Civ Eng* 22(10):4173–4184. <https://doi.org/10.1007/s12205-018-1347-1>
- Vapnik V (1998) *Statistical learning theory*. Wiley, New York
- Wang YW (1997) *IH: inducing model trees for predicting continuous classes*. Hamilton, New Zealand: University of Waikato, Department of Computer Science. <https://hdl.handle.net/10289/1183>