Predicting the infiltration characteristics for semi-arid regions using regression trees
Parveen Sihag, Munish Kumar and Saad Sh. Sammen

ABSTRACT
The study of the infiltration process is considered essential and necessary for all hydrology studies. Therefore, accurate predictions of infiltration characteristics are required to understand the behavior of the subsurface flow of water through the soil surface. The aim of the current study is to simulate and improve the prediction accuracy of the infiltration rate and cumulative infiltration of soil using regression tree methods. Experimental data recorded with a double ring infiltrometer for 17 different sites are used in this study. Three regression tree methods: random tree, random forest (RF) and M5 tree, are employed to model the infiltration characteristics using basic soil characteristics. The performance of the modelling approaches is compared in predicting the infiltration rate as well as cumulative infiltration, and the obtained results suggest that the performance of the RF model is better than the other applied models with coefficient of determination ($R^2$) = 0.97 and 0.97, root mean square error (RMSE) = 8.10 and 6.96 and mean absolute error (MAE) = 5.74 and 4.44 for infiltration rate and cumulative infiltration respectively. The RF model is used to represent the infiltration characteristics of the study area. Moreover, parametric sensitivity is adopted to study the significance of each input parameter in estimating the infiltration process. The results suggest that time ($t$) is the most influencing parameter in predicting the infiltration process using this data set.

Key words | cumulative infiltration, infiltration rate, M5 tree, random forest, random tree

HIGHLIGHTS
• Infiltration rate and cumulative infiltration characteristics in semi-arid region were modelled using regression tree.
• Random forest outperforms random tree as well as M5 tree to predict the infiltration characteristics.
• Time is observed as the most important parameter in affecting the prediction performance of both the infiltration rate and the cumulative infiltration of soil.

INTRODUCTION
Infiltration process influences the hydrological cycle, stream flow, groundwater recharge, irrigation, soil erosion and drainage system (Singh et al. 2018). Infiltration can be defined as irrigation or precipitation water moving into the soil (Sihag et al. 2018a, 2018b). The actual rate at which water moves into the soil is determined as infiltration rate (Haghighi et al. 2011). Infiltration process plays a significant role in the selection of crop type and agricultural practices. Prediction of infiltration is also necessary for the efficient design of irrigation systems (Hillel 1998). The sustainability of the
groundwater system also depends on this process (Chen & Young 2006).

The quantity and quality of recharged water is the most important parameter for water resources management especially in regions where water scarcity occurs. It divides water into the two most important hydrological segments, groundwater flow and surface flow. It is the main factor in watershed modelling for the prediction of surface runoff. Accurate prediction of infiltration rate is compulsory for reliable prediction of surface runoff (Diamond & Shanley 2005). At catchment level, infiltration characteristics are one of the main factors in determining the flooding condition (Bhave & Sreeja 2015). The significance of the infiltration process has required soil and water researchers to generate several models (e.g. Green & Ampt 1911; Kostiakov 1932; Horton 1944; Philip 1957; Holtan 1961; Soil Conservation Service 1972; Smith & Parlange 1978; Swartzendruber 1987; Singh & Yu 1990; Sihag et al. 2017a). However, only a few developed models have been effectively used with field data. Parameter calculation is the main criterion for selection of any model over other models. These infiltration models can be divided into three groups: physical models, semi-empirical models and empirical models.

The main problem for modelling the infiltration process is the variation of soil type and texture (Machiwal et al. 2006). Variation in the infiltration process in any watershed causes difficulty in management of water for agricultural practices. The process of infiltration is influenced by several aspects; including soil type, texture, soil hydraulic properties, soil moisture profile, and precipitation and climate properties. The spatial and temporal prediction of the infiltration process in the real environment cannot be at present deduced by direct or indirect methods. Infiltration models are based on the law of mass conservation and Darcy’s law (physical models), a few models rely on field and laboratory measurements (empirical models) and some models are based on simple hypotheses about the infiltration rate and cumulative infiltration relation quantities (semi-empirical models). These infiltration models are developed for vertically homogeneous soils with constant initial moisture content in the soil, constant density and over horizontal planes. A very few researchers have made attempts to use soft computing in the prediction of soil infiltration (Sy 2006; Sihag et al. 2018a, 2018b).

During the last few years, several researchers have used soft computing techniques such as adaptive neuro fuzzy inference system (ANFIS), artificial neural network (ANN), support vector machine (SVM), M5 tree, Gaussian process (GP), random forest (RF) and gene expression programming (GEP) in the prediction of hydraulic properties and infiltration process (Schaap & Leij 1998; Erzin et al. 2009; Anari et al. 2011; Das et al. 2011; Arshad et al. 2013; Al-Sulaiman 2015; Elbisy 2015; Esmaeinenjad et al. 2015; Al-Sulaiman et al. 2016; Rahmati 2017; Singh et al. 2017; Tiwari et al. 2017; Sihag et al. 2017b, 2017c; Sihag 2018) and found it to work equally well or better than empirical relations. A very few studies have been carried out on the spatial variability of soil hydraulic properties in arid regions. However, to the best of our information, no study has been carried out for describing the spatial variability of the soil water infiltration process using soft computing techniques in semi-arid regions of India. Thus, the main objective of this research is to develop infiltration models using regression-tree-based computing approaches and to determine the spatial variability at the same scale in some locations of Haryana (India). Both the cumulative infiltration as well as infiltration rate are predicted using random tree, random forest and M5 tree-based models. In the last few decades conventional infiltration models have been widely used for the calculation or estimation of infiltration rate and cumulative infiltration such as Kostiakov, Philip, Novel, Horton, Holton etc. These models are point- and location-specific. The aim of this investigation is to develop a general model for a complete study area using tree-based models.

STUDY AREA AND DATA COLLECTION

Four districts of Haryana State (Hisar, Kurukshetra, Kaithal, and Karnal) are selected for the infiltration measurements (Figure 1). The study area comprises a total of 17 sites represented by the numbers on the district maps (Figure 1).

The infiltration observations have been recorded in the field with the help of a double ring infiltrometer. The double ring infiltrometer has two concentric rings having diameter 30 and 60 cm for the inner and outer ring, respectively, with 30 cm depth. Both rings are driven about 10 cm deep into the soil by using a hammer without undue

Corrected Proof

by guest

disturbance to the soil surface. Both rings are filled with water to equal depth and the initial reading of the water level is noted. The rate of water level drop against time in the inner ring of the infiltrometer is noted at regular intervals until a constant rate is attained, which is called the steady-state infiltration rate. The soil samples for calculating the soil properties are collected from their respective locations selected for the measurement of infiltration rate. The sample is withdrawn along the side of the infiltrometer and represents the soil properties of that location. The properties of the soil samples are measured in the laboratory. The properties of soils respective to each location are listed in Table 1.

**DATA SET**

The data set used for the modelling in this study is a collection of infiltration observations yielded from the field experiments from 17 sites. In this study, infiltration rate ($I$) and cumulative infiltration ($Cu$) are considered as the

<table>
<thead>
<tr>
<th>Site no.</th>
<th>Sand (%)</th>
<th>Clay (%)</th>
<th>Silt (%)</th>
<th>Dry density (gm/cm$^3$)</th>
<th>Moisture content (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47.50</td>
<td>25.20</td>
<td>27.30</td>
<td>1.64</td>
<td>12.74</td>
</tr>
<tr>
<td>2</td>
<td>50.71</td>
<td>23.19</td>
<td>26.10</td>
<td>1.66</td>
<td>19.85</td>
</tr>
<tr>
<td>3</td>
<td>39.84</td>
<td>52.34</td>
<td>7.82</td>
<td>1.58</td>
<td>19.74</td>
</tr>
<tr>
<td>4</td>
<td>42.85</td>
<td>24.00</td>
<td>33.15</td>
<td>1.58</td>
<td>18.39</td>
</tr>
<tr>
<td>5</td>
<td>48.74</td>
<td>29.38</td>
<td>18.88</td>
<td>1.67</td>
<td>8.77</td>
</tr>
<tr>
<td>6</td>
<td>59.58</td>
<td>30.72</td>
<td>9.70</td>
<td>1.60</td>
<td>14.21</td>
</tr>
<tr>
<td>7</td>
<td>26.63</td>
<td>41.82</td>
<td>31.55</td>
<td>1.54</td>
<td>18.55</td>
</tr>
<tr>
<td>8</td>
<td>46.70</td>
<td>16.56</td>
<td>36.74</td>
<td>1.65</td>
<td>5.28</td>
</tr>
<tr>
<td>9</td>
<td>24.81</td>
<td>39.62</td>
<td>35.57</td>
<td>1.24</td>
<td>19.06</td>
</tr>
<tr>
<td>10</td>
<td>79.73</td>
<td>6.45</td>
<td>13.83</td>
<td>1.60</td>
<td>8.47</td>
</tr>
<tr>
<td>11</td>
<td>84.14</td>
<td>7.83</td>
<td>8.03</td>
<td>1.75</td>
<td>7.47</td>
</tr>
<tr>
<td>12</td>
<td>66.63</td>
<td>19.62</td>
<td>13.75</td>
<td>1.70</td>
<td>12.71</td>
</tr>
<tr>
<td>13</td>
<td>44.27</td>
<td>23.12</td>
<td>32.61</td>
<td>1.65</td>
<td>11.56</td>
</tr>
<tr>
<td>14</td>
<td>45.67</td>
<td>39.16</td>
<td>15.17</td>
<td>1.62</td>
<td>8.33</td>
</tr>
<tr>
<td>15</td>
<td>26.12</td>
<td>66.14</td>
<td>7.74</td>
<td>1.61</td>
<td>18.60</td>
</tr>
<tr>
<td>16</td>
<td>19.73</td>
<td>62.41</td>
<td>17.86</td>
<td>1.61</td>
<td>15.37</td>
</tr>
<tr>
<td>17</td>
<td>32.71</td>
<td>54.21</td>
<td>13.08</td>
<td>1.75</td>
<td>7.64</td>
</tr>
</tbody>
</table>
output variables whereas time ($t$), dry density ($\rho$), moisture content ($M$), and percentages of sand ($S_a$), silt ($S_i$), and clay ($C$) are the input variables. The basic soil properties respective to each site are measured from the laboratory experiments conducted on the field soil samples. The total data set is split in two groups: one used for model development (training data) and the other used for model validation (testing data). Training data involves 70% data chosen randomly from the complete data set, and the remaining 30% data forms the testing data. The features of the data set used for modelling the infiltration characteristics of the soil are presented in Table 2.

**MODELLING METHODS**

**Random tree (RT)**

Random trees are decision/regression-tree-based models used for classification/regression problems. Random model trees are essentially the combination of two existing algorithms: single model trees are combined with random forest ideas. Every leaf in model trees holds a linear model which is optimized for the local subspace described by that leaf (Pfahringer 2010). The collection of tree predictors in random trees is called a forest. With random features at each node, a random tree is a tree drawn randomly from a set of possible trees. The distribution of trees is uniform and each tree in the set of trees has an equal chance of being sampled. Random trees can be generated efficiently and the ensemble of random trees generally leads to accurate models (Zhao & Zhang 2008).

**Random forest (RF)**

The random forest proposed by Breiman (2001) is a classification and regression method. The performance of single decision trees is shown to be improved considerably by adding randomness to the training data using a random forest. The algorithm combines the concept of bagging (Breiman 1996) and the random selection feature (Amit & Geman 1997). In bagging, the training data is sampled with replacement for each tree. Therefore, a new training set (bootstrap samples) is drawn with replacement from the original training set. During the selection of the bootstrap samples some of the training data may be left out of the sample and some may be repeated in the sample (Adusumilli et al. 2013). To generate a tree, instead of computing the best split among attributes for each node, only a random subset of all attributes is chosen at every node, and the best split for that subset is computed.

The procedure for the random forest regression model is as follows:

1. Draw bootstrap samples from the data set.
2. Generate a regression tree with the following modifications: at each node, choose the best split from the randomly sampled $k$-tree predictors.
3. The models are tuned for the user-defined parameters: number of trees ($k$) grown and number of variables used at each node ($m$) to construct a tree.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min.</td>
<td>Max.</td>
</tr>
<tr>
<td>$t$ (minutes)</td>
<td>5.00</td>
<td>180.00</td>
</tr>
<tr>
<td>$S_a$ (%)</td>
<td>19.73</td>
<td>84.14</td>
</tr>
<tr>
<td>$C$ (%)</td>
<td>6.45</td>
<td>66.14</td>
</tr>
<tr>
<td>$S_i$ (%)</td>
<td>7.74</td>
<td>36.74</td>
</tr>
<tr>
<td>$\rho$ (gm/cm$^3$)</td>
<td>1.24</td>
<td>1.75</td>
</tr>
<tr>
<td>$M$ (%)</td>
<td>5.28</td>
<td>19.85</td>
</tr>
<tr>
<td>$I$ (mm/hr)</td>
<td>0.25</td>
<td>252.00</td>
</tr>
<tr>
<td>$Cu$ (mm)</td>
<td>1.50</td>
<td>383.00</td>
</tr>
</tbody>
</table>
M5 tree (M5)

Quinlan (1992) proposed this high-accuracy regression algorithm for dealing with continuous-class problems with high dimensionality. M5 generates a tree-based model which has multivariate linear regression models at the leaves. The first step is to build a tree, which is similar to inducing a regular decision tree. The difference is that the splitting criterion is based on maximizing the expected reduction in error instead of maximizing the information gain at each interior node. The standard deviation reduction (SDR) can be calculated by:

\[
SDR = sd(K) - \sum \frac{|K_i|}{|K|} sd(K_i)
\]  

(1)

where \( K \) represents number of instances that reach the node; \( K_i \) represents the number of instances that have the \( i^{th} \) outcome of the potential set; and \( sd \) represents the standard deviation of the target values of the instances in the data set \( K \).

The second step is to construct a multivariate linear regression function using the instances at each non-terminal node and then prune the built tree back from the leaves. Instead of using all attributes, this multivariate linear regression function only uses the attributes that are referenced by tests or linear functions somewhere in the subtree at this node.

The last step is to improve the prediction accuracy by applying some smoothing techniques. The smoothed predicted values of M5 are backed up from the leaf node to the root node (Li & Li 2011).

Performance assessment parameters

The performance of various modelling techniques has been carried out using various statistical performance assessment parameters, which are coefficient of determination \((R^2)\), root mean square error (RMSE) and mean absolute error (MAE) to check the statistical difference between the actual and the predicted values of the implemented soft computing approaches.

Coefficient of determination \((R^2)\): The coefficient of determination is used to measure of success of numeric prediction. The coefficient of determination \((R^2)\) is computed as follows (Adnan et al. 2013; Malik et al. 2019):

\[
R^2 = \frac{n \sum_{i=1}^{n} A_{obs_i} A_{pred_i} - \left( \sum_{i=1}^{n} A_{obs_i} \right) \left( \sum_{i=1}^{n} A_{pred_i} \right)}{\sqrt{n \left( \sum_{i=1}^{n} A_{obs_i}^2 \right) - \left( \sum_{i=1}^{n} A_{obs_i} \right)^2} \sqrt{n \left( \sum_{i=1}^{n} A_{pred_i}^2 \right) - \left( \sum_{i=1}^{n} A_{pred_i} \right)^2}}
\]

(2)

where \( A_{obs_i} \) is the observed values; \( A_{pred_i} \) is the predicted values; and \( n \) is the number of observations. The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of observed \((A_{obs})\) and predicted values \((A_{pred})\). The range of \( R^2 \) is 0 to 1.0. A correlation of 1.0 indicates that correlation is perfect among observed \((A_{obs})\) and predicted \((A_{pred})\) values, while a correlation of 0.0 indicates that there is no linear relationship among the observed \((A_{obs})\) and predicted \((A_{pred})\) values.

Mean absolute error (MAE)

The mean absolute the error is used to measure the success of numeric estimation. The mean absolute error (MAE) is computed as follows (Sammen et al. 2017; Pham et al. 2021):

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |A_{obs_i} - A_{pred_i}|
\]

(3)

MAE can range from 0 to \( \infty \). An efficiency of 0 (MAE = 0) corresponds to a perfect match of modelled \( A \) to the observed data.

Root mean square error (RMSE)

Mean-squared error (MSE) is the most commonly statistical evaluator for performance evaluation, and root mean square error is the square root of the mean-squared error after giving it the same dimensions as the predicted values themselves. This method exaggerates the prediction...
error minus the difference between prediction value and actual value. The root mean square error (RMSE) can be computed from the following equation (Sammen et al. 2020):

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (A_{\text{obs}} - A_{\text{pred}})^2}
\]  

(4)

RMSE can range from 0 to \(\infty\). An efficiency of 0 (RMSE = 0) corresponds to a perfect ideal model which predicts exact values as observed values.

**EVALUATION OF MODELS**

The performance measures adopted for model assessment are coefficient of determination \(R^2\), root mean square error (RMSE) and mean absolute error (MAE). The value of \(R^2\) determines the closeness of the regression line to the observed data. \(R^2 = 1\) indicates perfect fitting of the regression line to the observed data. RMSE and MAE are the error values that measure the accuracy of the models in approximating the actual observations. Lower values of RMSE and MAE indicate better prediction of the actual data. Two output values: infiltration rate and cumulative infiltration, are predicted relative to the input variables (Table 2) using the regression-tree-based models. In this study, WEKA 3.9 software is used for all applied regression-tree-based model development and validation. The models are developed using the training data set after setting the values of tuning parameters (user-defined) respective to each regression tree model. The performance of the constructed model is tested on the unseen testing data set using the statistical performance measures \(R^2\), RMSE and MAE. Several models are constructed and tested by assigning the values of the user-defined parameters to the modelling techniques and the best one is chosen based on optimal fitting of the model on the testing data set. The procedure is repeated for all the regression tree models using both outputs and the tuning parameters selected for performance analysis are shown in Table 3.

**RESULTS AND DISCUSSION**

**Modelling of infiltration rate**

The models are formed and checked for accuracy after setting the tuning parameters to near-optimal values (Table 3) in order to predict the values of soil infiltration rate. The simulation results of training and testing data are presented in Figures 2 and 3, respectively, which show the scattering of the data points for the prediction of three regression-tree-based models viz. random tree (RT), random forest (RF) and M5 tree. To investigate the superiority of the modelling techniques, the statistical performance measures acquired by the models with the observed infiltration rate data are elucidated in Table 4. As indicated from Figure 2, the scattering of training data predicted by the M5 tree with both pruned as well as unpruned model is higher relative to random tree and random forest. The higher values of infiltration rate

| Table 3 | Tuning parameters of regression models |
|---|---|---|
| Regression model | Infiltration rate prediction | Cumulative infiltration prediction |
| Random model tree | Number of randomly chosen attributes = 4 | Number of randomly chosen attributes = 3 |
| Random forest | \(k = 200, m = 1\) | \(k = 78, m = 2\) |
| M5 tree | Number of instances allowed at a leaf node = 5 | Number of instances allowed at a leaf node = 5 |

**Figure 2** | Scatter plot of infiltration rate prediction by the regression models using training data.
predicted by the M5 tree model scatter relatively more than the lower values and degrade the modelling performance. Statistical measures from the M5 tree confirm the poor performance of model development with the training data as the RMSE and MAE values are substantially higher. The performance of the model is improved with the testing data set, but the testing performance is inferior to the other regression tree models. The trained model based on the random tree is best \((R^2 = 0.99)\) as the data prediction is close to the line of perfect agreement in the training stage but the model is not capable of giving accurate predictions with the testing data as compared with the random forest model. The performance of the model developed with RT drops at the validation stage having RMSE and MAE values of 10.82 and 5.98, respectively (Table 4). Closer fitting of predicted points to the perfect agreement line using the RF model in the testing stage reveals its suitability in the generalization of the infiltration rate data as the error values (RMSE = 8.10 and MAE = 5.74) are minimum (Table 4). Figure 4 shows the residual error values obtained from the experimental observations and predictions of infiltration rate with regression tree models. The residuals are minimum with the RT model during the training stage, but the deviation of the model is higher from the zero error line during the testing stage relative to the RF model. M5 tree-based models have larger residuals during model training but the quality of the models improve during testing with smaller error values, but still, the models are ranked last after the RT model. Therefore, modelling results based on this study indicate superior performance of the RF model in simulating the actual infiltration rate of soil followed by the RT and M5 tree models, respectively.

### Modelling of cumulative infiltration

In this section, the cumulative infiltration of soil is simulated using random tree (RT), random forest (RF) and M5 tree (M5) models with the basic soil properties. The scattering plots of the regression-tree based models for the training and testing data sets are depicted in Figures 5 and 6, respectively. Table 5 provides the information about the statistical measures obtained from the observed training and testing cumulative infiltration data with their corresponding predicted values by the models. The scattering of the training data predicted by the M5 pruned and unpruned models is higher, particularly for the larger values, as compared with the RT and RF models (Figure 5). The predicted data of the RT and RF model fits closely to the perfect agreement line

### Table 4 | Performance of regression models in simulating the infiltration rate

<table>
<thead>
<tr>
<th>Regression model</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(R^2)</td>
<td>RMSE (mm/hr)</td>
</tr>
<tr>
<td>Random model tree</td>
<td>1.00</td>
<td>0.62</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.98</td>
<td>9.93</td>
</tr>
<tr>
<td>M5 tree pruned</td>
<td>0.81</td>
<td>26.81</td>
</tr>
<tr>
<td>M5 tree unpruned</td>
<td>0.81</td>
<td>26.81</td>
</tr>
</tbody>
</table>

### Table 5 | Performance of regression models in simulating the cumulative infiltration

<table>
<thead>
<tr>
<th>Regression model</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(R^2)</td>
<td>RMSE (mm)</td>
</tr>
<tr>
<td>Random model tree</td>
<td>0.99</td>
<td>0.92</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.99</td>
<td>10.33</td>
</tr>
<tr>
<td>M5 tree pruned</td>
<td>0.81</td>
<td>28.92</td>
</tr>
<tr>
<td>M5 tree unpruned</td>
<td>0.82</td>
<td>28.68</td>
</tr>
</tbody>
</table>
having lower values of RMSE and MAE than the M5 tree in the model development stage. The RT model perfectly predicts the training data with a maximum value of $R^2 = 0.99$ and minimum values of RMSE $= 0.92$ and MAE $= 0.54$ as most of the data points reside on the perfect agreement line, but the prediction capability of the model degrades during validation with the testing data. Analysing the testing data scatter plot (Figure 6), the prediction tendency of the RF model is superior as compared with the other regression models as the predicted points are comparatively closer to the perfect agreement line than the other models. The lowest values of RMSE $= 6.96$ and MAE $= 4.44$ obtained from the RF model at the testing stage affirm the stronger capability of the model in the generalization of the cumulative infiltration data (Table 5). So, statistical performance measures acquired from the testing data set suggest better performance by the RF model as compared with the RT and M5 tree models in approximating the cumulative infiltration of soil. Figure 7 shows the residual errors observed with the current data set using different regression models. The residual error values peak in the training stage, mostly predicted by the M5 tree model, while the RT model has smaller residual values. The residuals are minimum with the RF model relative to the RT model in the testing stage, which assures its higher predictability.

**SIMULATION OF INFILTRATION CHARACTERISTICS OF SOIL BY RANDOM FOREST MODEL**

Based on the modelling results discussed above, it is clear that the RF model has higher generalization performance in simulating the infiltration rate as well as cumulative
infiltration of soil as compared with the other applied models. In this section, infiltration characteristics experimentally observed from their respective sites are compared with the predicted values of the RF model. In order to accomplish this task, the RF model developed with the training data set is tested on the complete data set of the locations. Both outputs, infiltration rate and cumulative infiltration, observed from the field are presented with the RF model-based predictions for their respective sites (Figure 8). The model trained for both outputs using the tuning parameters (Table 3) is implemented to their respective data sets for prediction. For comparative analysis, the results are plotted for each site (specified by the location number). The infiltration rate/cumulative infiltration values are plotted against time to show the variation of the model in predicting the observed data. The results reveal that the prediction of points by the RF model agrees reasonably well with the observed infiltration values, particularly in predicting the infiltration rate, and the model has a decent tendency to follow the actual observations.

PARAMETRIC SENSITIVITY

The sensitivity of each input variable in predicting the output variables (infiltration rate and cumulative infiltration of soil) is tested. The random forest (RF) model developed with the training data set is used to study the parametric importance in prediction. Models are developed using various combinations of input parameters by removing each parameter one by one and the sensitivity is judged by means of $R^2$ and RMSE. The fluctuation in statistical measures with various input combinations accounts for the change in performance of each model in the prediction of both outputs. Analysing the combination from Table 6, it is revealed that the removal of time ($t$) from the input parameters causes higher fluctuation in both $R^2$ as well as RMSE. The combination drastically increases the RMSE value, which makes time ($t$) the most important variable among the other input variables in affecting the prediction of infiltration rate as well as cumulative infiltration. The absence of other input parameters in the RF model yields a small change in $R^2$ and RMSE values. Removal of the parameters sand, clay and silt, one by one, marginally decrease the RMSE and hence slightly improve the efficiency of the model in predicting the infiltration rate, while on the other hand, neglecting these parameters has a negative effect in cumulative infiltration prediction with slight increment in RMSE values (Table 6).

Infiltration is characterised as the flow of water into the subsurface from above ground. The topic of infiltration has received a great deal of attention because of its importance to topics as widely ranging as irrigation, contaminant transport, groundwater recharge, and ecosystem viability. In the last few decades conventional infiltration models have been widely used for the calculation or estimation of infiltration rate and cumulative infiltration, such as Kostiakov, Philip, Novel, Horton, Holton etc. Machiwal et al. (2006) observed that the infiltration process was well described by the Philip model in a wasteland of Kharagpur, India. However, soil management practices influencing the final infiltration rate are a major factor in deciding the applicability of these models.
Figure 8 | Comparison of observed and RF model predicted infiltration characteristics of soil respective to each site.
Thus, the variability of soil infiltration characteristics and goodness of fit of the infiltration models for different soils should be given due consideration in infiltration modelling studies for predicting the constant infiltration rate. The applicability of these models for estimating the infiltration rate with different soil management has been examined by researchers. Gifford (1976) observed that amongst the Horton, Kostiakov and Philip models, the Horton model was the best model to fit the infiltration data in mostly semi-arid regions from Australia, but under specific conditions only. Mishra et al. (2003) compared the performance of 14 different conventional infiltration models using 243 sets of infiltration data collected from field and laboratory tests conducted in India and the USA on soils ranging from coarse sand to fine clay. These models are point- or location-specific, and until now no general model has been developed. The results of this study suggest that the RF-based model is suitable for predicting the infiltration characteristics of the study area. The RF model is a general model for the study area and gives closer values to the observed values. The RF model is used to represent the infiltration characteristics of the study area and it saves time and effort in comparison with experimentation and conventional models.

### CONCLUSIONS

Prediction of infiltration characteristics is essential for several hydrologic studies associated with surface and subsurface flow. In this study, the effectiveness of regression-tree-based models in simulating the infiltration characteristics of soil is evaluated. In estimating the infiltration rate and cumulative infiltration of soil, random forest outperforms random tree as well as M5 tree with $R^2 = 0.97$ and 0.97, RMSE = 8.10 and 6.96 and MAE = 5.74 and 4.44, respectively. The training and testing results of the random forest encourage the utility of this method relative to the other tested methods and show a good potential in representing the infiltration characteristics of the study area. Based on the sensitivity results, time is observed as the most important parameter in affecting the prediction performance of both the infiltration rate and the cumulative infiltration of soil.

Finally, it is worth noting that the proposed approach uses only the six fundamental parameters of $t$, $\rho$, $M$, $S_a$, $S_i$ and $C$ for prediction of the infiltration rate and cumulative infiltration of soil. This is of importance for situations where there is no possibility of measuring other parameters like porosity and lower depth soil type. Considering the ability of the RF model in handling missing values in any data sets as well as its capacity in bypassing noisy data, the implemented soft computing methodology showed the potential to be accurately used in approximation of the infiltration process. The RF model’s ability needs to be compared with hybrid techniques to try to develop a more accurate model. More experimental data is also required for better and general results.

### DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

### REFERENCES


Chen, L. & Young, M. H. 2021 Distribution of natural radionuclides in the surface soil in some areas of agriculture and grazing located in west of Riyadh, Saudi Arabia. Journal of Applied Life Sciences International 7 (2), 1–12.

Chen, L. & Young, M. H. 2013 Predicting saturated hydraulic conductivity by artificial intelligence and regression models. ISRN Soil Science 2013, 308159.


Sammen, S. S., Ghorbani, M. A., Malik, A., Tikhamarine, Y., AmirRahmani, M., Al-Ansari, N. & Chau, K.-W. 2020 Enhanced artificial neural network with harris hawks optimization for predicting scour depth downstream of


First received 23 November 2020; accepted in revised form 29 January 2021. Available online 18 February 2021.