Machine learning-based modeling of saturated hydraulic conductivity in soils of tropical semi-arid zone of India

SATISH BHAURAO MORE[1], PARESH CHANDRA DEKA[2], AMIT PRAKASH PATIL[3,*] and SUJAY RAGHAVENDRA NAGANNA[4]

[1] Department of Civil Engineering, N. K. Orchid College of Engineering & Technology, Solapur 413002, Maharashtra, India
[2] Department of Water Resources and Ocean Engineering, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India
[3] Department of Civil Engineering, Annasaheb Dange College of Engineering and Technology, Ashta 416301, Maharashtra, India
[4] Department of Civil Engineering, Siddaganga Institute of Technology, Tumakuru 572103, Karnataka, India
e-mail: moresb23@gmail.com; dekanitk@gmail.com; amitpatil.nitk@gmail.com; sujay.gopan@gmail.com

Abstract. Saturated hydraulic conductivity (K_{fs}) is the major parameter that affects the movement of water and solutes in soil strata. Although one can estimate the K_{fs} directly by using various field or laboratory methods, they turn out to be more time-consuming and painstaking while characterizing the spatial variability of K_{fs}. For this reason, some recent researches employ indirect approaches such as pedotransfer functions (PTF) and surface modeling methods for estimating K_{fs} of several scales. Pedotransfer functions are often developed by relating the K_{fs} with readily available soil properties such as bulk density, porosity, sand content, silt content, and organic material. The present research explores the suitability of Extreme Learning Machine (ELM) in developing PTF's for K_{fs} by using basic soil properties. In-situ field tests and laboratory experiments on collected samples were performed to acquire the datasets necessary for the analysis. Three competitive soft computing approaches, namely the ELM, Support Vector Machine (SVM), and Adaptive Neuro-Fuzzy Inference System (ANFIS) based on Fuzzy C-means Clustering optimized by Genetic Algorithm were exercised for developing the K_{fs} models. Further, the performance of these approaches in modeling K_{fs} was evaluated using various statistical mertics. The performance of ELM was found to be good in comparison to the other two models, with sufficiently good NSE values. The ELM model provided K_{fs} predictions at the Murarji Peth and Punanaka sites with an NSE of 0.90 and 0.83, respectively, while at the Mulegoan site, the ANFIS model was better with R = 0.80 and NSE = 0.64.

Keywords. Saturated hydraulic conductivity; Guelph permeameter; ELM; SVM; ANFIS.

1. Introduction

Hydraulic conductivity is an essential and predominant soil property for perceiving and replicating various soil-water interaction processes such as infiltration, contaminant plume in the vadose zone, and subsurface drainage [1, 2]. Appropriate knowledge of field-scale soil hydraulic conductivity is also crucial for estimating primary consolidation settlements of foundation [3], seepage loss from earthen dams [4], drainage filter design [5], and assessing bearing capacity for ground improvement [6]. It is also a coordinating factor influencing root zone processes and soil surface macro-porosity in irrigated fields. The hydraulic conductivity of natural field soils govern the variations in groundwater residence and travel time (inflow and outflow rates), modifying or altering the quantity of subsurface flow to the nearby streams or underlying aquifer [7, 8].

Several anthropogenic factors influence the soil's physical, chemical, and biological properties in a significant way. One such anthropogenic impact includes changes in land use and land cover, which are the most dynamic phenomenon influenced by land use for various purposes, directed by cultural, social, and economic interactions. Changes in land use result from human demands arising from changed natural, economic or geo-political issues. The consequences are either modification or conversion from one land-use type to another. Land-use changes largely influence the hydraulic conductivity as it is responsible for

altering the pore space geometries of the soil [8]. In contrast to land that is covered with vegetation, barren terrain has a lower saturated hydraulic conductivity [9]. Impervious surfaces (such as roads, footpaths) created as a result of urbanization either by ramming or consolidating the landscape affect the hydraulic conductivity of soil and cause flooding [10].

Soil horizons or terrain is another morphometric factor that influences local soil properties significantly. Sloping terrains are commonly seen in the Indian subcontinent. Researchers have reported that slope influences hydraulic properties such as moisture distribution, infiltration rate, cumulative infiltration, and hydraulic conductivity. Hydraulic conductivity in the hill slope is a deciding factor for slope stability. Having the knowledge of hydraulic conductivity is essential for landslide and soil stability analysis [11].

Saturated hydraulic conductivity is a scale-dependent parameter [12, 13]; hence, many observations are required to characterize a given site. Both in-situ and laboratory methods are troublesome, laborious, and tedious. Also, the outcomes may not be precise because of inherent variability (spatial and temporal) in soil hydraulic properties. All these factors have prompted the evolution and boundless use of indirect methods (pedotransfer functions PTF) that estimate K_{fs} from easily, universally available, and routinely measured basic soil properties such as percentage of sand, silt, clay, bulk density, and organic matter within the soil volume [14–16]. The development of PTFs have helped in the widespread application of models for simulating various soil properties.

The researchers developed and used several PTFs to model K_{fs} of soils [17]. The strategies run from simple lookup tables to complex data-driven techniques like regression analysis [18], neural networks [19, 20], Support vector machines [21, 22], Adaptive neuro-fuzzy interface system [23], group method of data handling, and regression trees [24]. Lookup tables contain tabular relations between soil textural soil class and related hydraulic properties [25, 26]. Complex methods like regression analysis investigate relations between dependent and independent variables for developing PTF's [27, 28]. The artificial neural networks (ANN), which use the computational model of the human brain and pattern recognition approach to map the input and output data relations, were found effective in modeling soil hydraulic properties [29, 30]. Few researchers developed hybrid structures such as neuro-fuzzy systems that take advantage of both ANN and Fuzzy logic to model field-scale soil hydraulic conductivity [31, 32]. The group method of data handling (GMDH), which works by generating analytical function in a feedforward multilayer neural network (FFMLNN) based on a quadratic node transfer function, is also being recently used by several researchers [33, 34] for developing PTFs. The ANN models

minimize prediction errors by implementing the empirical risk minimization principle, due to which there exists a possibility of solutions trapping into local minima. The SVM, which operates on the structural risk minimization principle, is known to overcome this disadvantage. Several SVM-based PTFs performed superior to conventional regression and ANN-based PTFs while estimating the soil K_{fs} [35, 36]. Other techniques like regression trees have also been used by many researchers [37] for developing PTFs. Regression-based PTFs were used extensively in the past owing to their simplicity. However, the PTFs based on the pattern recognition approach (neural networks) have been popular in the last decade. The widely used soft computing based PTFs, model soil hydraulic properties without considering the physics involved in the processes. However, their inherent ability to adapt to complex input-output relationships produces sufficiently better estimates of the soil parameter of interest. ELM, introduced by Huang *et al* [38], has gained the attention of researchers owing to its properties like quick learning, superior generalization capability, inherent competence to set its internal parameters, and robustness. The ELM based PTFs to model soil hydraulic properties (specifically soil hydraulic conductivity) is rarely available in the literature.

As tropics enfold approximately 40% of the earth's land surface and since the tropical semi-arid regions worldwide are becoming agriculturally less productive, the edaphological issues of soils of tropical environments need to be considered for sustainable management protocols [39]. Hence, modeling the field saturated soil conductivity (K_{fs}) of tropical semi-arid soils supports the rational management of soil properties to get a better yield of crops from agriculture. Modeled K_{fs} data could also benefit in modeling water transport and waste contaminant migration through the soil. This research intends to explore the suitability of ELM in the development of PTF's for modeling K_{fs} by using basic soil properties such as bulk density, porosity, and soil texture as model parameters. Further, the ELM-based PTF model performance is compared with that of SVM and ANFIS based PTFs in modeling field-scale saturated hydraulic conductivity (K_{fs}).

2. Theoretical overview

2.1 *Extreme learning machine*

It was observed that a single-layered feedforward system (SLFN) could be effectively trained with randomly adopted input weights [40, 41], leading to the development of extreme learning machines. ELM belongs to the category of a single hidden neural network trained with the Moore-

Penrose generalized inverse learning algorithm, which adjusts the weights of the output layer analytically even though the input weights and hidden biases are chosen randomly. The basics of the ELM algorithm are very briefly given below.

For N arbitrary distinct inputs (x_i, y_i) with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$; a standard SLFN with N hidden nodes and activation function $\mathfrak{f}$ can be modeled as

$$\sum_{i=1}^{\mathcal{N}} \beta_i \mathfrak{f}(w_i x_j + b_i), \quad j \in \{1, 2, 3, \ldots, N\}, \qquad (1)$$

where w_i are the input weights to the ith neuron in the hidden layer, b_i the biases, and β_i are the output weights [42–44]. It has advantages like extremely fast training speed, no over-fitting problem, and good generalization performance [45]. ELM trains the SLFN in the following stages; firstly, it maps the feature randomly and then solves it for linear parameters. This strategy makes the ELM algorithm work faster than the SLFN algorithm.

2.2 *Support vector machine*

SVM is a robust and reliable regression paradigm that performs excellently even with a limited amount of data. SVM is based on statistical learning theory, which refers to a vast set of mathematical implementations for understanding data [46]. The SVM includes a statistical framework that avoids posterior computing probabilities while building a decision machine from a dataset. The model is based on support vectors which represents a vector subset of data taken from the training set. Following equation is balanced form of the considered support vectors.

$$f(\overline{x}) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(\overline{x}, \overline{x}^j) + b \qquad (2)$$

where α_i^* and α_i are the Lagrange multipliers and the expression $K(\overline{x}, \overline{x}^j)$ is the kernel function. SVM formulates a solution using a quadratic optimization problem with linear inequality constraints. When used for regression, the SVM leads to a globally optimal solution and has high generalization ability. In any kernel-based learning method, choosing a suitable kernel is crucial for achieving good performance. More importantly, when the learning method deals with multiple heterogeneous data sources, it must consider multiple kernels [47, 48]. The approaches such as artificial neural networks minimize the empirical risks, whereas SVM uses a structural risk minimization (SRM) principle, which aims at minimizing an upper bound on the generalization error. Thus, SVM has higher prediction capabilities on the unseen dataset [49–52].

2.3 *ANFIS based on Fuzzy C-means Clustering optimized by Genetic Algorithm*

ANFIS was proposed by Jang *et al* [53], wherein the fuzzy if-then rules (between input and outputs) in a fuzzy network are constructed using neural networks. Hence, ANFIS can be considered as a fuzzy inference system equipped with the advantages of neural networks. An ANFIS model combines the transparent and linguistic representation of a fuzzy system with the learning ability of ANN. Fuzzy C-means (FCM) is a data clustering technique in which a dataset is grouped into different clusters. The datasets are grouped such that every data point belongs in a cluster to a certain degree. For example, if a particular point lies at the center of a cluster, it will have a high degree of membership or belonging to that cluster. Whereas, if it lies far away from the cluster's center, the degree of membership will be low to that cluster. Dunn, [54] initially developed this FCM method which marks the mean location of each cluster based on initial guesses. Then the centers are moved to suitable locations by iteratively updating the cluster centers and membership grades. The iteration is based on minimizing an objective function presented in [55] that represents the distance from any given data point to a cluster center weighted by that data point's membership grade. To optimize the weighting exponent in the FCM, Genetic Algorithm (GA) was used in this study.

Genetic algorithm (GA) is a stochastic search iterative method based on the evolutionary theory. It is generally used to optimize nonlinear problems and is based on Darwin's evolution theory [56]. Conceptually Genetic Algorithm implements genetic operators like crossover, mutation, and selection for up-gradation and searches for the best population by artificially imitating the natural evolution process. The genetic algorithm is initiated with an initial population of possible solutions, called individuals. The evolution theory aims to compute its fitness function and then determine whether an individual can enter the next generation. The fitness value F of an individual can be expressed as:

$$Fi = k \left(\sum_{i-1}^{n} abs(P_i - O_i) \right) \qquad (3)$$

where P_i is the predicted value, and O_i is the desired value of an individual. Then selected individuals are then manipulated using crossover and mutation. This process is continued till the GA search is converged or the termination criterion is satisfied. This idea is based on the theory that better parents may probabilistically generate better offspring.

3. Study area and data analysis

Vertisol soils occupy 22% of the total geographic area of India (figure 1). These soils are present in western (Maharashtra state) and central (Madhya Pradesh) parts of the country. Vertisols are soils with a high content of clay minerals that shrink and swell with the change in water content. The clay minerals adsorb water and increase in volume (swell) when wet and then shrink as they dry, forming large, deep cracks. Measuring the hydraulic conductivity of Vertisols is a difficult task, as the pore size changes with moisture content. In the present study, the saturated hydraulic conductivity of Vertisol soil was determined through Guelph Permeameter from three diverse soil profiles within the study area – Solapur, Maharashtra state, India, shown in figure 2 geographically located at 17.65° N Latitude and 75.90° E Longitude. The Solapur district is located at 483 m above mean sea level, and the mean annual rainfall is around 723 mm (highest 1292 mm and lowest 270 mm). The rainfall is scanty, erratic and ill-distributed, due to which water scarcity conditions prevail in this place. May is the hottest, and December is the coldest month of the year. In general, the climate of the study area can be classified as "semi-arid."

The soil in this area is derived from basalt and is underlain by partially disintegrated rock locally called 'murum.' The three sites selected for this study enfold three diverse soil profiles; the Mulegoan site was an agricultural land during a fallow period; the Punanaka site was barren land with sloping terrain, and the Murarji Peth site was a pastureland (grass cover) with mild downward slope followed by undulated terrain.

A sampling at all these sites was done in a 0.76 ha area (190 m × 40 m). Every site was divided into small grids of 10 m × 10 m, as shown in figure 3. In-situ saturated hydraulic conductivity (K_{fs}) was determined by using Guelph permeameter (Model: 2800K1). The K_{fs} measurements were taken from every grid point at a depth of 15 cm, 30 cm, and 45 cm. The Guelph measurement was used for estimating a quasi-steady discharge Q of water infiltrating into a vertical borehole of radius a (3 cm) in which the water level is maintained at a constant height H (5 cm/10 cm) above the bottom of the borehole. The steady-state water level change (rates) in the Guelph reservoir is established to find the saturated hydraulic conductivity of the soil under investigation [57]. Every in-situ K_{fs} measurement by Guelph permeameter took over two hours of time and overall about 100 K_{fs} measurements were made at
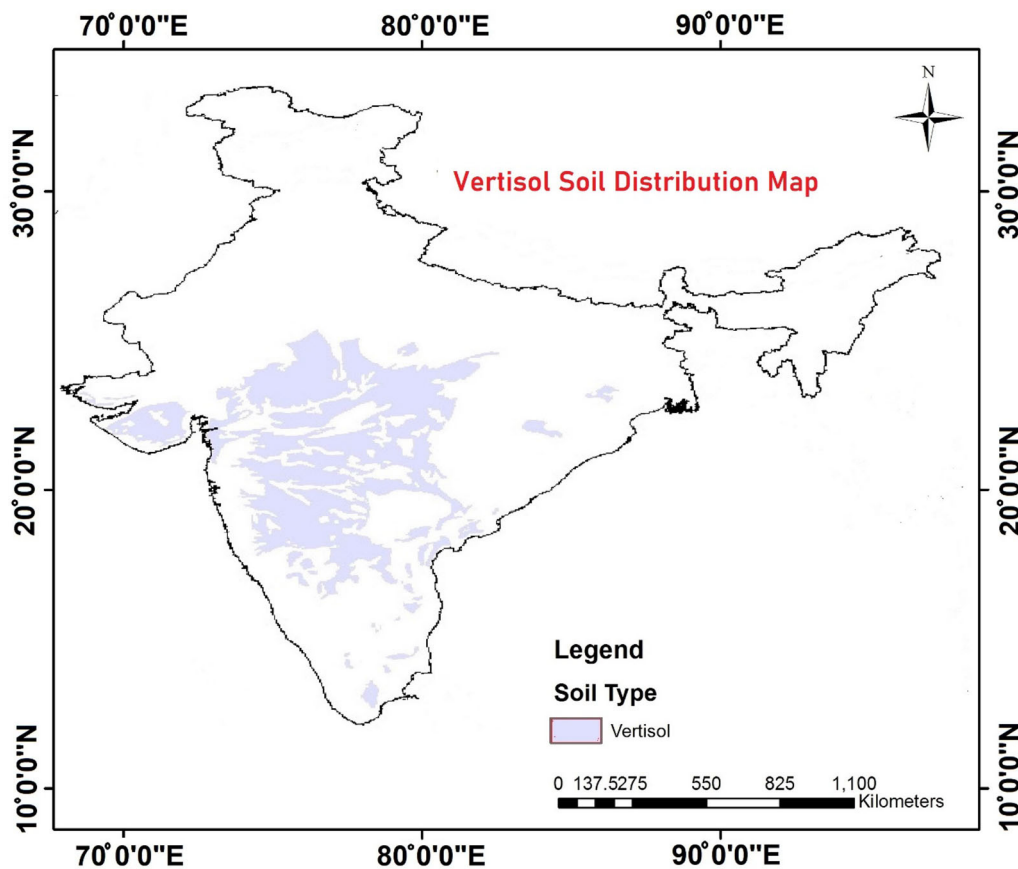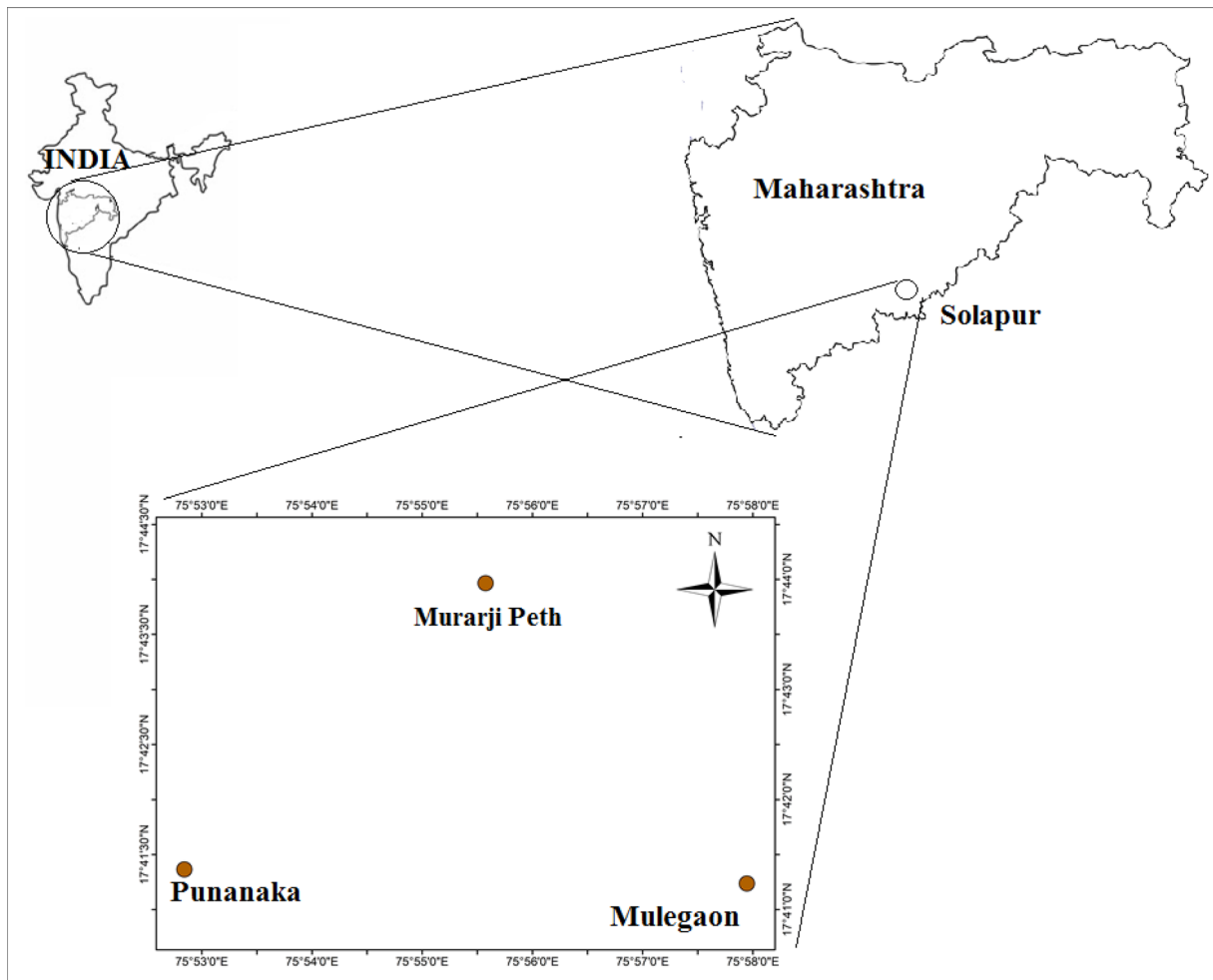

Figure 1. Vertisol soil distribution in India.

Figure 2. Location of Study area.

each depth at all three locations. Thus, in total, 300 K_{fs} measurements were gathered from each site location. The summary of raw dataset is reported in highlight Appendix 1 to give a sense of values and its ranges.

Soil samples were collected from all three depths after sampling K_{fs} using a permeameter to carry out laboratory analysis [58]. All the soil samples were collected carefully using two cylindrical cores (100 mm diameter and 125 mm depth). Soil samples of the first cylindrical core were used to determine dry bulk density, particle density [59], and organic matter content. Samples of the second core were used to determine soil texture using hydrometer analysis.

3.1 *Data preprocessing*

The K_{fs} dataset was tested for the normal distribution by using statistical techniques. From the quantile-quantile (Q-Q) plot, it was found that they were not normally distributed. To get normally distributed data, they were log-transformed. An example of the Q-Q plot before and after data transformation is presented in figure 4. A summary of statistics of soil parameters sampled at all three sites and all three depths (15 cm, 30 cm, and 45 cm) is presented in table 1. Each dataset was normalized between 0.05 and 0.95 using min-max normalization to allocate initial weights based on the distribution of the parameter rather than its magnitude. The output obtained by the modeling technique was then denormalized for comparison with the original data.

4. Methodology

4.1 *Selection of input parameters*

Based on the literature, it was observed that factors like bulk density (BD), porosity (n), sand % (S), silt % (Si), clay % (C), particle density (G), and organic matter (OM) affect K_{fs} [14–16]. While dealing with multivariate regression, multicollinearity of independent variables is a general problem that the researchers face. This occurs when two or
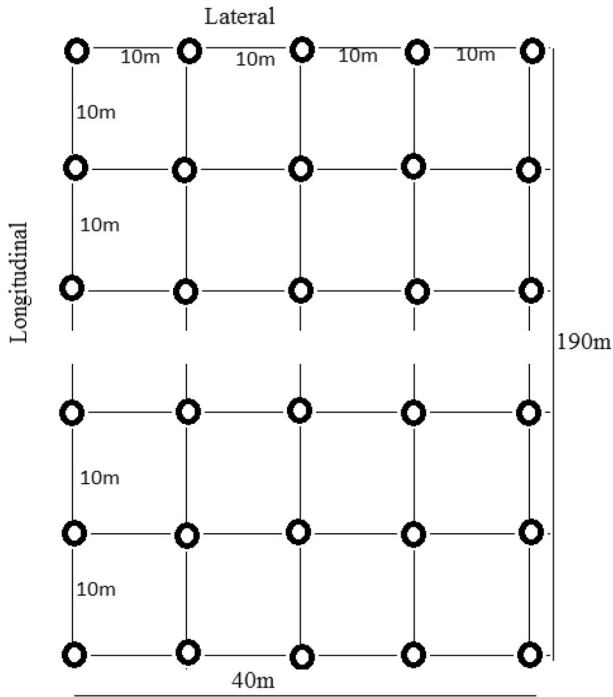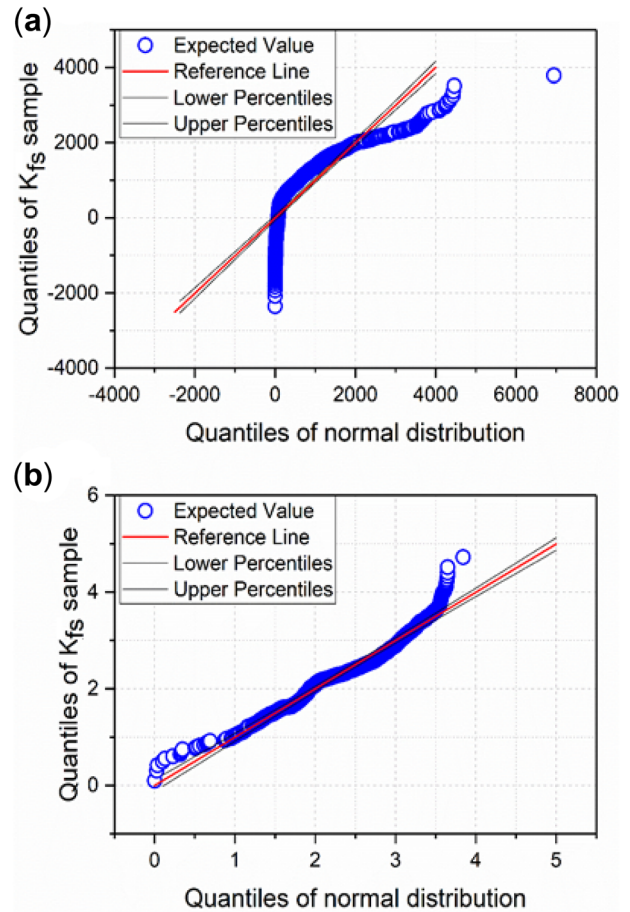
Figure 3. Sampling scheme grid at Murarji Peth, Mulegaon and Punanaka sites.

Figure 4. Q-Q Plot for saturated hydraulic conductivity. (a) Before transformation and (b) After transformation.

more independent variables in a regression model are moderately or significantly correlated. Multicollinearity affects only the coefficients and P-values; and does not influence the accuracy of the model predictions, and the goodness-of-fit statistics. Statistician Jim Frost [60] believes that if the primary objective is to make only predictions, and there is no need to understand the role of each independent variable, then it is not mandatory to reduce severe multicollinearity. Hence, in the present case study, amongst all the parameters, the most influential parameters were determined by using step-wise regression having significance at $P < 0.01$. The input parameters selected for modeling K_{fs} of various sites are presented in table 2.

4.2 Model development

The division of the data set into training and testing sets is an essential component of AI model development. Various models were tested for determining the best distribution of data sets into training and testing data sets. Several random training and testing data partitions (70:30, 67:33, and 90:10) were experimented to find the best proportion for training the models. The optimal data proportion for training was 70% for models of Mulegaon site, 67% for models of Punanaka site, and 90% for the models of Murarji Peth site. While solving the optimal model complexity, the machine learning models behave differently from one case to another under varied data divisions. Hence, the attained data division of datasets is very normal considering the optimal performance of the developed models.

The ELM model consisted of input, hidden, and output layers, as shown in figure 5. The input parameters used are presented in table 2. The output layer had one neuron representing the estimated K_{fs}. In the hidden layer, a maximum of 100 neurons was tested for each model. The number of neurons in the hidden layer was varied between 2 and 100, with Radial Basis Function as the kernel for all ELM models tested.

The SVM model developed includes the optimization of its hyper-parameters. The Radial Basis Function (RBF) kernel was used during the development of SVM models. During the training stage, hyperparameters: C (cost function), kernel width (γ), and epsilon-insensitive loss function (ε) were optimized by using a grid search optimizer. As these parameters are interdependent, the grid search operation was performed in two stages. During the first stage, a coarse grid was used (keeping a wide range for the parameters with considerable increment). Furthermore, a refined grid was used to determine the optimum values of C, γ, and ε. The combination of hyperparameters, which

Table 1. Summary of statistics of soil parameters (log Transformed) sampled at all three sites and all three depths (15 cm, 30 cm and 45 cm).

		Mean	SD	Kurtosis	Skewness	Minimum	Maximum
MP		2.067	0.804	−0.505	−0.231	0.002	3.638
MU	K_{fs} (m yr^{-1})	2.576	0.621	−1.073	−0.167	0.506	3.647
PN		2.610	0.598	−0.575	−0.330	1.098	3.842
MP		1.622	0.073	−0.954	0.030	1.504	1.761
MU	Porosity (%)	1.668	0.053	−0.962	0.174	1.504	1.773
PN		1.699	0.055	−0.243	−0.292	1.517	1.786
MP		1.265	0.319	−0.453	0.458	0.699	1.944
MU	% Sand	1.394	0.281	−0.600	0.636	0.778	1.944
PN		1.492	0.296	−1.043	0.000	0.845	1.940
MP		1.472	0.389	2.108	−1.296	0.000	1.924
MU	% Clay	1.407	0.392	0.494	−0.904	0.000	1.914
PN		1.311	0.417	0.034	−0.690	0.000	1.914
MP		1.461	0.305	−0.495	−0.513	0.301	1.903
MU	% Silt	1.440	0.313	1.448	−0.935	0.000	1.863
PN		1.415	0.306	1.473	−0.969	0.000	1.863
MP		3.152	0.044	−0.619	−0.785	3.053	3.203
MU	BD (kg m^{-3})	3.128	0.044	−1.443	−0.162	3.047	3.202
PN		3.099	0.040	−0.690	0.263	3.042	3.201
MP		0.923	0.287	20.751	−3.684	−1.495	1.174
MU	OM (g kg^{-1})	0.981	0.118	6.767	−1.589	0.273	1.144
PN		0.391	0.227	1.002	−1.085	−0.541	0.712
MP		0.396	0.015	−1.194	−0.362	0.370	0.427
MU	G (g/cm^3)	0.406	0.010	0.493	−0.155	0.369	0.437
PN		0.408	0.017	−1.007	0.561	0.372	0.452

The values of all soil parameters are log transformed, MP—Murarji Peth site, MU—Mulegaon site & PN—Punanaka

Table 2. Input parameters selected based on step-wise regression for various dataset.

Sl. No.	Dataset	Input parameters
1	Murarji Peth	Porosity (n), Sand % (S), clay % and bulk density (BD).
2	Mulegaon	Porosity (n), Sand % (S), clay % and bulk density (BD).
3	Punanaka	Porosity (n), Sand % (S), Silt % (Si) and bulk density (BD)

resulted in a minimum mean square error value, was taken as optimal hyperparameters. To avoid overfitting, a four-fold cross-validation approach was used during the training phase.

In the Fuzzy C-means clustering-based ANFIS model, the number of clusters was determined based on a trial-and-error approach. The optimal number of clusters were searched in the range of 2–20 with a step size of one. Genetic algorithm was employed to optimize the weighting exponent (m). A larger 'm' value is known for making FCM more robust to noise as well as outliers. However, larger 'm' values greater than the theoretical upper bound will make the sample mean a unique optimizer. Hence, genetic algorithm was employed to search the optimal value of 'm', which is neither less than one and greater than the theoretical upper bound value of the sample data [61].

4.3 *Model evaluation criteria*

The following performance metrics were employed to evaluate the model efficiencies.

i. Coefficient of correlation (R)

$$R = \frac{\sum_{i=1}^{N}\left(O_i - O_{avg}\right)\left(P_i - P_{avg}\right)}{\sqrt{\sum_{i=1}^{N}\left(\left(O_i - O_{avg}\right)\right)^2}\sqrt{\sum_{i=1}^{N}\left(P_i - P_{avg}\right)^2}} \quad (4)$$

ii. Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{1}^{N}(O - P)^2}{N}} \quad (5)$$

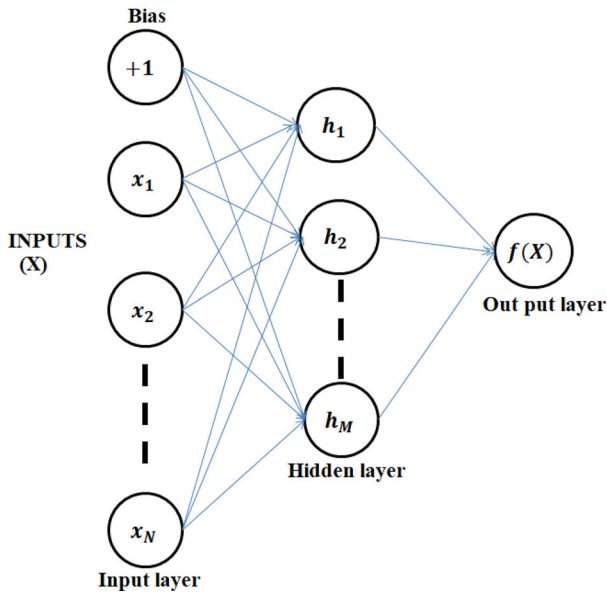and Normalized root mean square error (NRMSE) is

Figure 5. Architecture of ELM model.

$$NRMSE = \frac{RMSE}{O_{max} - O_{min}} \qquad (6)$$

iii. Mean relative error (MRE)

$$MRE = \frac{1}{N} \sum_{1}^{N} \left(\frac{O - P}{O} \right) \qquad (7)$$

iv. Nash–Sutcliffe model efficiency coefficient (NSE)

$$NSE = 1 - \left[\frac{\sum_{1}^{N}(O - P)^2}{\sum_{1}^{N}\left(O - O_{avg}\right)^2} \right] \qquad (8)$$

where O—the observed value of the variable, P—the predicted value of the variable, O_{max}—maximum value of the observed variable, O_{min}—minimum value of the observed variable, O_{avg}—average value of the observed variable, and P_{avg}—average value of the predicted variable, N—the number of observations.

5. Results and discussion

The results obtained from the field experimentation and various models developed have been analyzed and discussed here. In the first part, the statistical analyses of the soil dataset at each of the three locations are discussed, and the second part presents a performance analysis of the models developed for each station.

5.1 *Statistical analysis*

The soil parameters tested in the laboratory and the field include saturated hydraulic conductivity (K_{fs}), bulk density (BD), porosity (n), sand % (S), silt % (Si), clay % (C), particle density (G), and organic matter (OM). The statistical indices such as mean value, standard deviation, the minimum and maximum value of each soil parameter are presented in table 1. The maximum value of log transformed saturated hydraulic conductivity (lnK_{fs}) was observed at Punanaka (3.842 m/yr), and the minimum value was at the Murarji Peth site (0.002 m/yr). The standard deviation of lnK_{fs} was more at the Murarji Peth site (0.804) than the other two sampling stations. The correlation coefficient of various soil parameters with the saturated hydraulic conductivity (logarithmic terms) is presented in table 3. It could be observed that porosity had a more significant influence on K_{fs} with a correlation coefficient of 0.9 compared to other parameters. The next factor controlling K_{fs} was bulk density; it was negatively correlated with the K_{fs}, thereby showing a contrasting response. The soil water content differences intensify with decreasing bulk density. The other parameters, OM, and G held negligible correlation; and the parameters S, C, and Si were moderately correlated with the K_{fs} at all three locations, namely Murarji Peth, Mulegoan, and Punanaka. Previous studies have shown reduced K_{fs} values in soils with higher bulk density [62]. At the Murarji Peth site, the bulk density of soil was higher and minimum value of K_{fs} were observed at this location. This may be due to soil compaction by the cattle movement (grazing). Considering the barren land at the Punanaka site, two different soil mass classes were found at the top and bottom of the terrain. This may be due to the sloping topography of the site. At the higher elevations of the terrain, high K_{fs} values were observed because of the coarse-grained soil profile; and at the bottom stretch of the terrain, the soil with less K_{fs} value was found due to the accumulation of silt and other fine organics deposited by runoff.

5.2 *Performance evaluation of models*

Performances of all machine learning models developed were assessed using various statistical metrics as mentioned in section 4.3. The performance metrics of the models in both training and testing phases of all three locations are presented in tables 4, 5 and 6. Scatter and box plots of ELM, SVM, and ANFIS model predictions of the test phase are shown in figures 6, 7 and 8.

5.2.1 *Murarji peth site:* The ELM model provided K_{fs} predictions with an NSE = 0.90 in the testing phase. The performance of ELM was found to be exceptionally good in

Table 3. Correlation Coefficient of various soil parameters with saturated hydraulic Conductivity sampled at Murarji Peth (MP), Mulegaon (MU) and Punanaka (PN).

	K_{fs} (m yr^{-1})	Porosity %	Sand %	Clay %	Silt %	BD (kg m^{-3})	OM (g kg^{-1})	G
MP	1.00	0.90	0.76	−0.43	−0.14	−0.87	0.28	0.73
MU	1.00	0.88	0.58	−0.42	−0.28	−0.90	−0.05	0.23
PN	1.00	0.70	0.73	−0.59	−0.36	−0.71	0.10	0.58

All soil parameters are in logarithmic terms, MP – Murarji Peth site, MU – Mulegaon site & PN – Punanaka site

comparison with the other two models. The error statistic, MRE of the ELM model in the training phase was 0.02 ($\sim$0) and 0.18 in the testing phase (table 4). In terms of NRMSE also, the performance of ELM was excellent during both the training and the testing phase. The correlation coefficient (R) of SVM and ANFIS models was relatively good during the training phase. However, in the testing phase, both the models underperformed in comparison to the ELM model. The scatterplot of observed v/s predicted K_{fs} values of the testing phase is shown in figure 6. The predicted K_{fs} values by the ELM method were closer to the 1:1 line when compared to other two methods. The predictions by the SVM method during testing were below the 1:1 line indicating the underfitting of the values. Even though the performance of the ANFIS model was better than that of SVM, the predictions are unsatisfactory. The box plots of observed and predicted K_{fs} values by all three models are depicted in figure 6. The median of observed K_{fs} predicted by all three models is close to the first quartile, and most of the points are in between the lower quartile and median. More points were concentrated towards the first quartile, which implies that the predictions from all models were virtuous/superior when the magnitude of K_{fs} was low, but for higher values of K_{fs} prediction, the capability of the ELM model was relatively decent. The ELM also predicted the outliers with reliable accuracy, and the data distribution was similar to that of observed K_{fs}. The spread of data predicted by ANFIS was reasonably agreeable with that of observed K_{fs}; however, the prediction of the SVM model was poor, and most of the data points were within the third quartile, its upper whisker is very short in comparison with that of other two models.

5.2.2 *Mulegaon site:* In terms of all performance statistics, the performance of ANFIS was found to be satisfactory compared to the SVM and ELM models, as depicted in table 5, during both the training and testing phase. The NSE values of SVM and ELM models were relatively less than that of ANFIS predictions during the testing phase. Although the NRMSE of ANFIS and ELM predictions were equal, there exists a significant difference in MRE values of both models. The ANFIS model provided generalized predictions during the testing

Table 4. Performance of models for the data of Murarji Peth site (90% training data and 10% testing data).

|	Training			Testing		
Model	ELM	SVM	ANFIS	ELM	SVM	ANFIS
R	0.97	0.77	0.78	**0.96**	0.52	0.58
MRE	0.02	0.07	0.06	**0.18**	0.60	0.58
RMSE (m yr^{-1})	73.01	257.12	217.52	**219.03**	771.37	652.55
NRMSE	0.02	0.07	0.06	**0.06**	0.21	0.18
NSE	0.95	0.56	0.60	**0.90**	0.18	0.15

Table 5. Performance of models for the data of Mulegaon site (70% training data and 30% testing data).

	Training			Testing		
Model	ELM	SVM	ANFIS	ELM	SVM	ANFIS
R	0.94	0.80	0.87	**0.88**	0.74	0.80
MRE	0.05	2.29	1.08	**0.88**	1.60	0.63
RMSE (m yr^{-1})	332.23	978.60	495.34	**411.91**	631.70	378.04
NRMSE	0.07	0.22	0.11	**0.09**	0.14	0.09
NSE	0.87	0.48	0.72	**0.58**	0.34	0.64

Table 6. Performance of models for the data of Punanaka site (67% Training data and 33% testing data).

	Training			Testing		
Model	ELM	SVM	ANFIS	ELM	SVM	ANFIS
R	0.99	0.85	0.85	**0.94**	0.63	0.78
MRE	0.07	1.91	0.90	**0.71**	1.41	1.44
RMSE (m yr^{-1})	161.99	627.76	460.74	**464.83**	904.75	733.03
NRMSE	0.02	0.09	0.07	**0.07**	0.13	0.11
NSE	0.97	0.59	0.78	**0.83**	0.37	0.59

phase compared to the other two models. Figure 7 presents the scatterplot of observed v/s predicted K_{fs} values. The majority of the predicted K_{fs} values by the
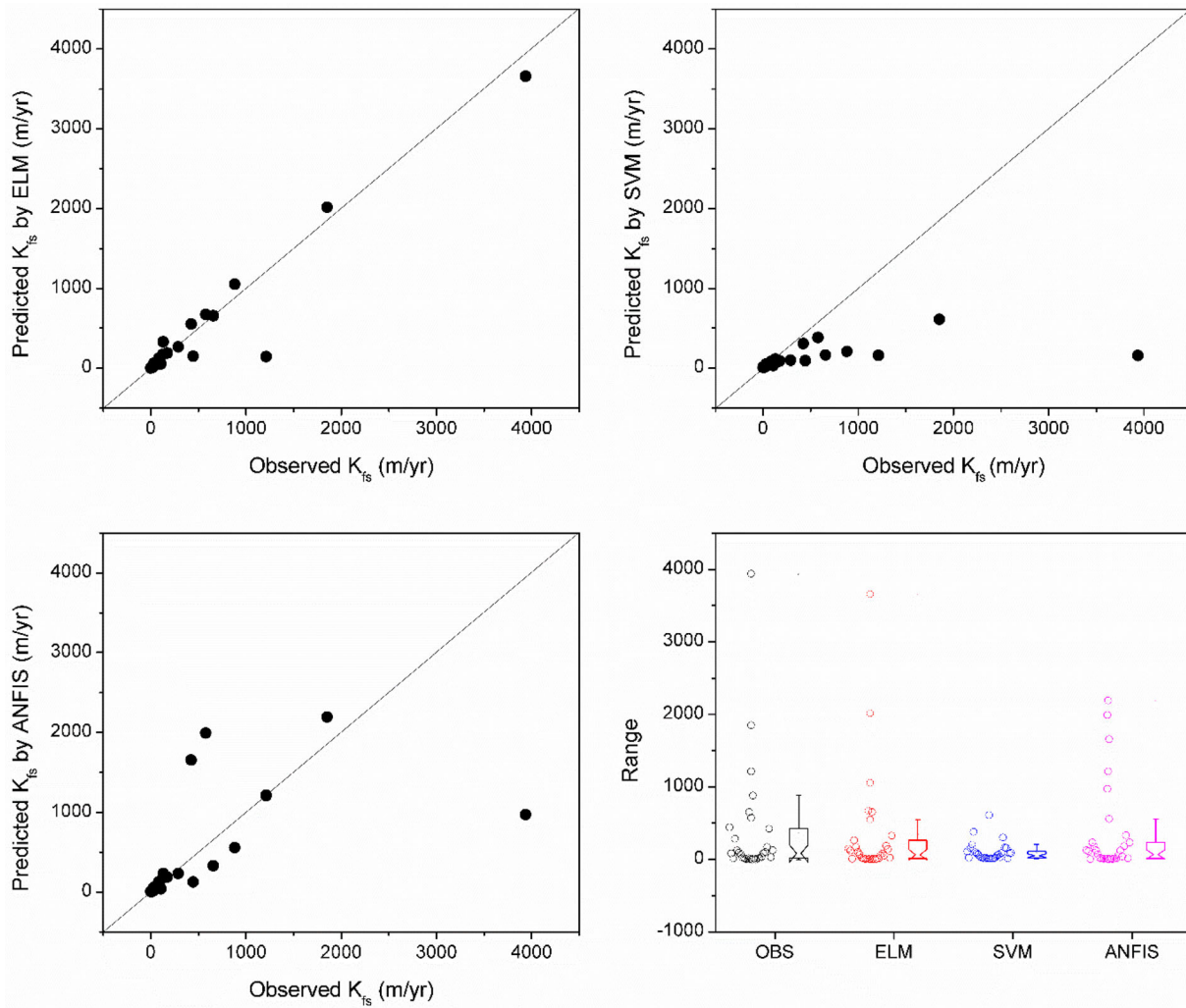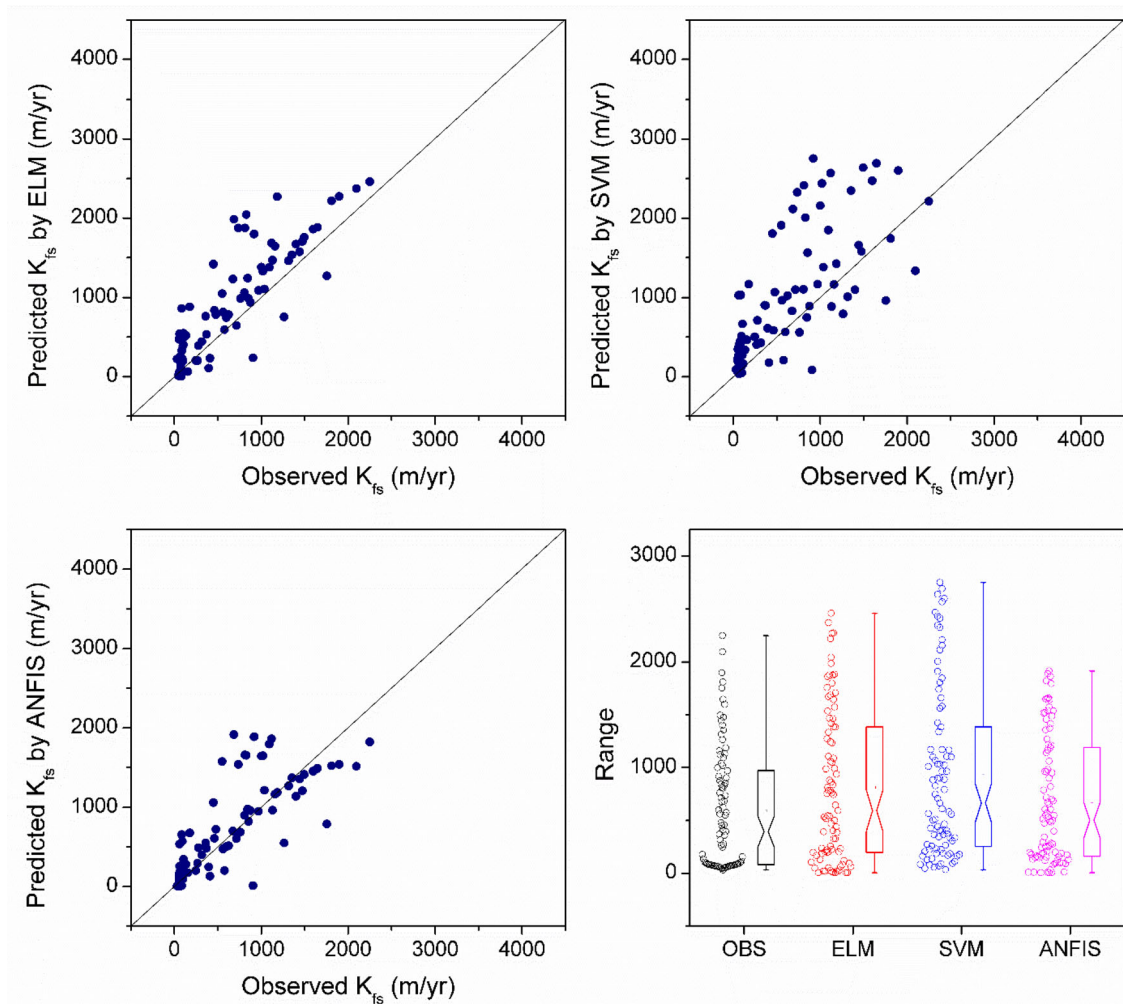
Figure 6. Scatter plot and Box plot of ELM, SVM and ANFIS models in testing of Murarji Peth site.

ANFIS model were close to the 1:1 line compared to the other two methods. The predictions by the SVM model were not satisfactory (over predicted). A bulk of ELM model predictions are scattered above 1:1 line signposting the over-prediction of K_{fs} values. The ANFIS model better predicted the smaller values of K_{fs}. Box plots for observed and predicted K_{fs} values by various methods are shown in figure 7. The median of observed K_{fs} was found lower than the median of predicted K_{fs} by all three models. The lower whisker of the observed K_{fs} is small, indicating lower 25% values of K_{fs} were closely spaced (variability in lower 25% value is relatively less). Upper whiskers of observed K_{fs} and predicted K_{fs} is large, indicating a more significant variation in K_{fs} values in the last (upper) quartile. Predictions of the ANFIS model were relatively good since its box plot approximately matches with the box

plot of observed K_{fs}. This may be because of the sensitivity of influencing parameters considered for modeling the hydraulic conductivity. The agricultural activities that cannot be accounted for during modeling may play an important role in developing a good data driven model for modelling hydraulic conductivity.

5.2.3 *Punanaka site:* During the training and testing phase, the NSE of the ELM model was found to be 0.97 and 0.83, respectively, which is a sign of good model performance (table 6). The ANFIS model held a satisfactory NSE=0.78 during the training phase; however, its performance during testing (NSE = 0.62) was not at par with that of the training phase. The ELM model had lower RMSE (about 48.62% and 36.58%) compared to SVM and ANFIS models, respectively. Barring a few outliers, all the predicted K_{fs} values of the
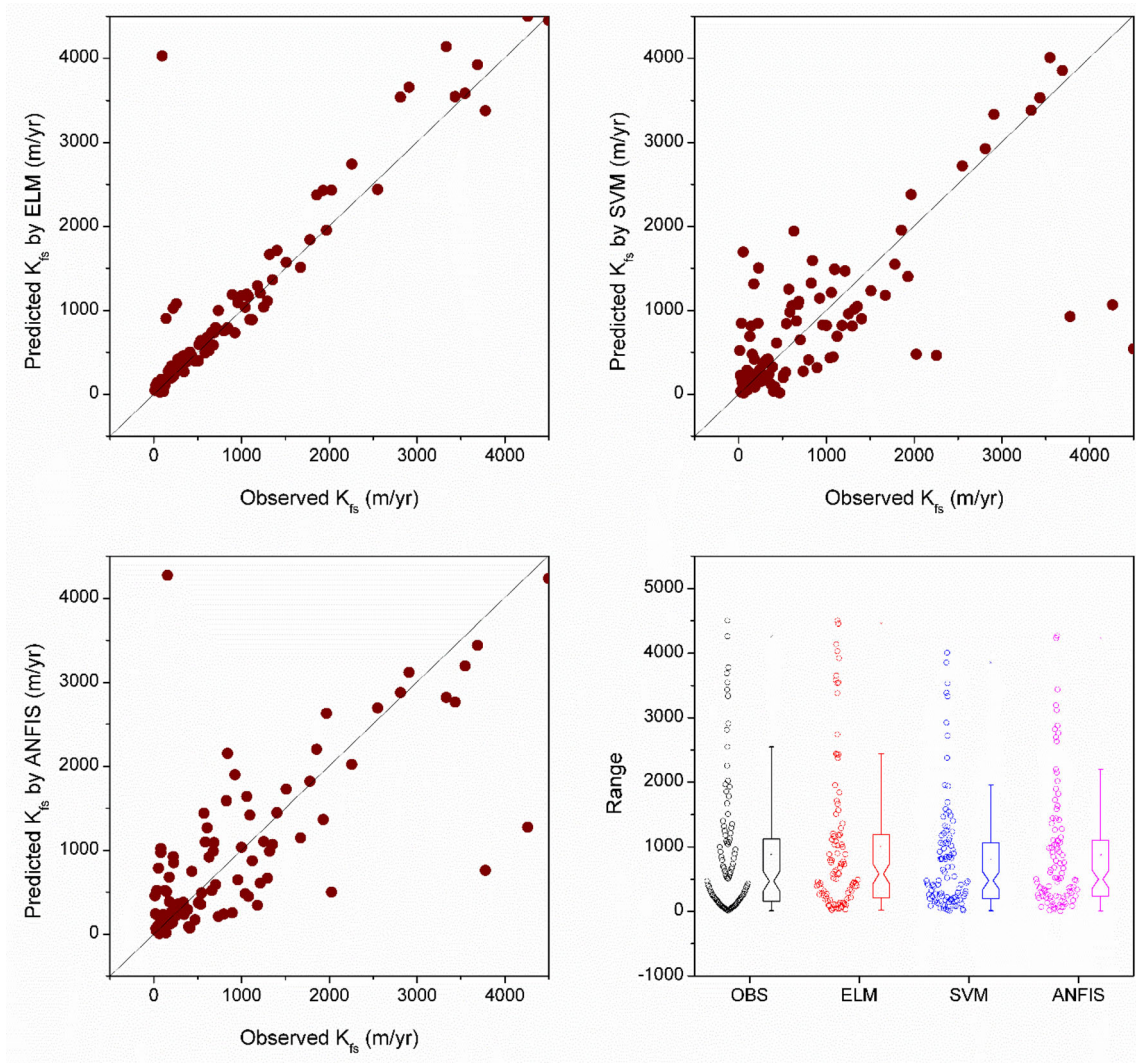
Figure 7. Scatter plot and Box plot of ELM, SVM and ANFIS models in testing of Mulegaon site.

ELM model of the testing phase (figure 8) are closely spread along the 1:1 line. SVM was found to overestimate the K_{fs} values, whereas the ANFIS prediction was found closely spread around the 1:1 line on either side of the line, but for higher K_{fs} values, it was found underestimating. The Box plots of observed K_{fs} against predicted K_{fs} values by ELM, SVM, and ANFIS, are depicted in figure 8. Median of observed and predicted K_{fs} values by the ELM model is approximately the same; the length of lower whisker, upper whisker, the distance between median and quartiles (lower quartile and upper quartile) is almost similar, implying that the spread of data is identical. However, the predictions are either on the higher or, the lower side of the observed values.

Lim and Kolay [63] predicted the hydraulic conductivity of tropical soils using different ANN architectures. They used bulk density, moisture content, dry density, void ratio, liquid limit, plastic limit, gravel, sand, silt, and clay percentages as input parameters and could predict hydraulic conductivity with a model efficiency NSE = 0.72 during the test phase. In another study by Arshad [64], the data from 175 soil profiles from different parts of Khuzestan province were modeled using ANFIS, ANN, and multiple linear regression (MLR) models. They obtained model efficiencies (Coefficient of determination) R^2= 0.71, 0.66, and 0.5, respectively, during the test phase. They used percentages of clay, silt, sand, and bulk density as model input parameters. In the present case, the ELM model provided an efficiency NSE=0.90 (Murarji Peth site) during the test phase while predicting the saturated hydraulic conductivity of the same tropical soils. The obtained modeling results portray ELM as a reliable learning machine to model highly variable and non-stationary data, as demonstrated in the current study. In general, the length of data used for training the model has a considerable effect on the accuracy of the predictions.

Figure 8. Scatter plot and Box plot of ELM, SVM and ANFIS models in testing of Punanaka site.

6. Conclusions

The current study aimed to predict the saturated hydraulic conductivity of tropical semi-arid zone soils using three well-recognized machine learning paradigms. Based on the analysis, it was found that porosity, sand content, bulk density, and clay/silt content were the influential parameters to model the saturated hydraulic conductivity of the soil. Porosity and sand content have shown a positive influence, and the bulk density and clay content have shown a negative correlation with the saturated hydraulic conductivity of the soil. A maximum value of 3.842 m/year (saturated hydraulic conductivity in the logarithmic scale) was observed at the Punanaka site, and a minimum value of 0.002 m/year was observed at the Murarji Peth site. The performance of the ELM model was found to be better than that of SVM and ANFIS models at Murarji Peth and Punanaka sites. The performance of the ELM model was not up to the expectation at the Mulegaon site, and the possible reason for this could be any extrinsic influencing parameters which are not accounted for, such as macro-hole distribution due to roots and biological activities that exist in agricultural lands. The multicollinearity of inputs could be one more reason for reduced efficiencies and has been reported as a limitation of the current study. The sensitivity of every parameter on the output variable is recommended to be reflected in future research.

Appendix 1

Appendix 1. Summary of statistics of soil parameters (raw data) sampled at all three sites and all three depths (15 cm, 30 cm and 45 cm).

		Mean	SD	Kurtosis	Skewness	Minimum	Maximum
MP		438.621	715.048	7.991	2.649	1.005	4347.959
MU	K_{fs} (m yr^{-1})	855.932	1004.590	2.377	1.677	3.205	4433.412
PN		870.332	1059.269	4.995	2.071	12.526	6942.543
MP		42.466	7.120	−0.779	0.301	31.900	57.700
MU	Porosity (%)	46.900	5.815	−1.086	0.341	31.900	59.300
PN		50.354	6.294	−0.809	−0.028	32.900	61.100
MP		24.617	21.280	1.923	1.722	5.000	88.000
MU	% Sand	31.117	23.445	0.372	1.360	6.000	88.000
PN		38.713	25.295	−1.063	0.696	7.000	87.000
MP		39.603	24.510	−1.302	0.234	1.000	84.000
MU	% Clay	34.817	23.076	−1.228	0.366	1.000	82.000
PN		29.633	22.012	−0.909	0.620	1.000	82.000
MP		35.780	20.978	−1.189	0.340	2.000	80.000
MU	% Silt	34.067	19.559	−1.267	0.311	1.000	73.000
PN		31.653	17.703	−1.045	0.342	1.000	73.000
MP		1427.575	137.403	−0.779	−0.669	1130.634	1595.306
MU	BD (kg m^{-3})	1350.064	136.395	−1.477	−0.080	1114.520	1593.557
PN		1262.860	118.143	−0.349	0.439	1100.454	1590.217
MP		9.502	3.408	0.069	−0.706	0.032	14.945
MU	OM (g kg^{-1})	9.879	2.340	−0.374	−0.197	1.874	13.947
PN		2.758	1.143	−0.963	−0.105	0.288	5.154
MP		2.491	0.088	−1.189	−0.326	2.342	2.673
MU	G (g/cm^3)	2.546	0.057	0.433	−0.074	2.340	2.738
PN		2.558	0.103	−0.940	0.616	2.355	2.829

Abbreviations

K_{fs}	Saturated Hydraulic Conductivity
lnK_{fs}	Log Transformed Saturated Hydraulic Conductivity values
PTF	Pedotransfer Functions
ELM	Extreme Learning Machine
SVM	Support Vector Machine
ANFIS	Adaptive Neuro-Fuzzy Inference System
GMDH	Group Method of Data Handling
FFMLNN	Feedforward Multilayer Neural Network
SLFN	Single-Layered Feedforward Network
SRM	Structural Risk Minimization
FCM	Fuzzy C-means
GA	Genetic Algorithm
Q-Q	Quantile-Quantile
BD	Bulk Density
n	Porosity
S	Sand %
Si	Silt %
C	Clay %
G	Particle Density
OM	Organic Matter
RBF	Radial Basis Function
R	Coefficient of Correlation
RMSE	Root Mean Square Error
NRMSE	Normalized Root Mean Square Error
MRE	Mean Relative Error
NSE	Nash–Sutcliffe Efficiency

Declarations

Conflict of interest: All the authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

[1] Buttle J M and House D A 1997 Spatial variability of saturated hydraulic conductivity in shallow macroporous soils in a forested basin. *J. Hydrol.* 203(1–4): 127–142

[2] Mallants D, Mohanty B P, Vervoort A and Feyen J 1997 Spatial analysis of saturated hydraulic conductivity in a soil with macropores. *Soil Technol.* 10(2): 115–131

[3] Karim M R and Lo S C 2015 Estimation of the hydraulic conductivity of soils improved with vertical drains. *Comput. Geotech.* 63: 299–305

[4] Himanshu N and Burman A 2019 Seepage and stability analysis of durgawati earthen dam: a case study. *Indian Geotech. J.* 49(1): 70–89

[5] Canga E, Iversen B V and Kjaergaard C 2014 A simplified transfer function for estimating saturated hydraulic conductivity of porous drainage filters. *Water Air Soil Pollut.* 225(1): 1–13

[6] Wong C, Pedrotti M, El Mountassir G and Lunn R J 2018 A study on the mechanical interaction between soil and colloidal silica gel for ground improvement. *Eng. Geol.* 243: 84–100

[7] Jiang X W, Wan L, Wang X S, Ge S and Liu J 2009 Effect of exponential decay in hydraulic conductivity with depth on regional groundwater flow. *Geophys. Res. Lett.* 36(24)

[8] Naganna S R, Deka P C, Ch S and Hansen W F 2017 Factors influencing streambed hydraulic conductivity and their implications on stream–aquifer interaction: a conceptual review. *Environ. Sci. Pollut. Res.* 24(32): 24765–24789

[9] Tian J, Zhang B, He C and Yang L 2017 Variability in soil hydraulic conductivity and soil hydrological response under different land covers in the mountainous area of the Heihe River Watershed, Northwest China. *Land Degrad. Dev.* 28(4): 1437–1449

[10] Gori A, Blessing R, Juan A, Brody S and Bedient P 2019 Characterizing urbanization impacts on floodplain through integrated land use, hydrologic, and hydraulic modeling. *J. Hydrol.* 568: 82–95

[11] Malaya C and Sreedeep S 2013 A study on unsaturated hydraulic conductivity of hill soil of north-east India. *ISH J. Hydraul. Eng.* 19(3): 276–281

[12] Seyfried M S and Wilcox B P 1995 Scale and the nature of spatial variability: Field examples having implications for hydrologic modeling. *Water Resour. Res.* 31(1): 173–184

[13] Sobieraj J A, Elsenbeer H and Cameron G 2004 Scale dependency in spatial patterns of saturated hydraulic conductivity. *Catena* 55(1): 49–77

[14] Vereecken H, Schnepf A, Hopmans J W, Javaux M, Or D, Roose T, Vanderborght J, Young M H, Amelung W, Aitkenhead M and Allison S D 2016 Modeling soil processes: Review, key challenges, and new perspectives. *Vadose Zone J.* 15(5)

[15] Assouline S and Or D 2013 Conceptual and parametric representation of soil hydraulic properties: A review. *Vadose Zone J.* 12(4)

[16] Chapuis R P 2012 Predicting the saturated hydraulic conductivity of soils: a review. *Bull. Eng. Geol. Environ.* 71(3): 401–434

[17] Araya S N and Ghezzehei T A 2019 Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. *Water Resour. Res.* 55(7): 5715–5737

[18] Kotlar A M, Iversen B V and de Jong van Lier Q 2019 Evaluation of parametric and nonparametric machine-learning techniques for prediction of saturated and near-saturated hydraulic conductivity. *Vadose Zone J.* 18(1): 1–3

[19] Kashani M H, Ghorbani M A, Shahabi M, Naganna S R and Diop L 2020 Multiple AI model integration strategy—application to saturated hydraulic conductivity prediction from easily available soil properties. *Soil Tillage Res.* 196: 104449

[20] Williams C G and Ojuri O O 2021 Predictive modelling of soils' hydraulic conductivity using artificial neural network and multiple linear regression. *SN Appl. Sci.* 3(2): 1–3

[21] Sihag P, Singh V P, Angelaki A, Kumar V, Sepahvand A and Golia E 2019 Modelling of infiltration using artificial intelligence techniques in semi-arid Iran. *Hydrol. Sci. J.* 64(13): 1647–1658

[22] Naganna S R and Deka P C 2019 Artificial intelligence approaches for spatial modeling of streambed hydraulic conductivity. *Acta Geophys.* 67(3): 891–903

[23] Sihag P, Esmaeilbeiki F, Singh B, Ebtehaj I and Bonakdari H 2019 Modeling unsaturated hydraulic conductivity by hybrid soft computing techniques. *Soft Comput.* 23(23): 12897–12910

[24] Sihag P, Karimi S M and Angelaki A 2019 Random forest, M5P and regression analysis to estimate the field unsaturated hydraulic conductivity. *Appl. Water Sci.* 9(5): 1–9

[25] Carsel R F and Parrish R S 1988 Developing joint probability distributions of soil water retention characteristics. *Water Resour. Res.* 24(5): 755–769

[26] Agyare W A, Park S J and Vlek P L 2007 Artificial neural network estimation of saturated hydraulic conductivity. *Vadose Zone J.* 6(2): 423–431

[27] Jabro J D 1992 Estimation of saturated hydraulic conductivity of soils from particle size distribution and bulk density data. *Trans. ASAE* 35(2): 557–560

[28] Aimrun W and Amin M S 2009 Pedo-transfer function for saturated hydraulic conductivity of lowland paddy soils. *Paddy Water Environ.* 7(3): 217–225

[29] Al-Sulaiman M A and Aboukarima A M 2016 Prediction of unsaturated hydraulic conductivity of agricultural soils using artificial neural network and c. *J. Agric. Ecol. Res. Int.* 1–5

[30] Parasuraman K, Elshorbagy A and Si B C 2006 Estimating saturated hydraulic conductivity in spatially variable fields using neural network ensembles. *Soil Sci. Soc. Am. J.* 70(6): 1851–1859

[31] Al-Sulaiman M A 2015 Applying of an adaptive neuro fuzzy inference system for prediction of unsaturated soil hydraulic conductivity. *Biosci. Biotechnol. Res. Asia* 12(3): 2261–2272

[32] More S B and Deka P C 2018 Estimation of saturated hydraulic conductivity using fuzzy neural network in a semi-arid basin scale for murum soils of India. *ISH J. Hydraul. Eng.* 24(2): 140–146

[33] Nemes A T and Rawls W J 2004 Soil texture and particle-size distribution as input to estimate soil hydraulic properties. *Dev. Soil Sci.* 30: 47–70

[34] Pachepsky Y A and Rawls W J 1999 Accuracy and reliability of pedotransfer functions as affected by grouping soils. *Soil Sci. Soc. Am. J.* 63(6): 1748–1757

[35] Twarakavi N K, Šimůnek J and Schaap M G 2009 Development of pedotransfer functions for estimation of

soil hydraulic parameters using support vector machines. *Soil Sci. Soc. Am. J.* 73(5): 1443–1452

[36] Elbisy M S 2015 Support vector machine and regression analysis to predict the field hydraulic conductivity of sandy soil. *KSCE J. Civ. Eng.* 19(7): 2307–2316

[37] Mady A Y and Shein E V 2018 Support vector machine and nonlinear regression methods for estimating saturated hydraulic conductivity. Moscow University. *Soil Sci. Bull.* 73(3): 129–133

[38] Huang G B, Zhu Q Y and Siew C K 2004 Extreme learning machine: a new learning scheme of feedforward neural networks. *IEEE International Joint Conference on Neural Networks* (IEEE Cat. No. 04CH37541) 2004 Jul 25 (2: 985-990)

[39] Food F A 2016 Agriculture: key to achieving the 2030 agenda for sustainable development. Rome: *Food and Agriculture Organization of the United Nations.*:1-32 [Online]. Available: http://www.fao.org/publications/card/en/c/d569c955-8237-42bf-813e-5adf0c4241b9

[40] Huang G B, Zhu Q Y and Siew C K 2006 Extreme learning machine: theory and applications. *Neurocomputing.* 70(1–3): 489–501

[41] Tamura S I and Tateishi M 1997 Capabilities of a four-layered feedforward neural network: four layers versus three. *IEEE Trans. Neural Netw.* 8(2): 251–255

[42] Huang G B, Wang D H and Lan Y 2011 Extreme learning machines: a survey. *Int. J. Mach. Learn. Cybern.* 2(2): 107–122

[43] Huang G B 2014 An insight into extreme learning machines: random neurons, random features and kernels. *Cognit. Comput.* 6(3): 376–390

[44] Huang G, Huang G B, Song S and You K 2015 Trends in extreme learning machines: A review. *Neural Netw.* 61: 32–48

[45] Patil A P and Deka P C 2016 An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Comput. Electron. Agric.* 121: 385–392

[46] Vapnik V, Golowich S E and Smola A 1997 Support vector method for function approximation, regression estimation, and signal processing. *Adv. Neural Inf. Process. Syst.* 2: 281–287

[47] Vapnik V 1998 Statistical learning theory. Wiley, New York

[48] Vapnik V 1999 The nature of statistical learning theory. Springer, New York

[49] Raghvendra S and Deka P C 2014 Support vector machine applications in the field of hydrology: a review. *Appl. Soft Comput.* 19: 372–386

[50] Vapnik V N 1999 An overview of statistical learning theory. *IEEE Trans. Neural Netw.* 10(5): 988–999

[51] Cristianini N and Shawe-Taylor J 2000 An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge. p 208

[52] Schölkopf B, Smola A J, Williamson R C and Bartlett P L 2000 New support vector algorithms. *Neural Comput.* 12(5): 1207–1245

[53] Jang J S 1993 ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cybern.* 23(3): 665–685

[54] Dunn J C 1973 A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.* 3(3): 32–57

[55] Zhang Y, Wang W, Zhang X and Li Y 2008 A cluster validity index for fuzzy clustering. *Inf. Sci.* 178(4): 1205–1218

[56] Goldberg D E 1989 Genetic algorithms in search optimization and machine learning

[57] Moisture S 2012 Guelph permeameter: operating instructions. Soil Moisture Equipment Corporation, Santa Barbara

[58] IS 2720 (Various Parts) Methods of test for soils, Bureau of Indian Standards

[59] Santra P and Das B S 2008 Pedotransfer functions for soil hydraulic properties developed from a hilly watershed of Eastern India. *Geoderma* 146(3–4): 439–448

[60] Frost J 2019 *Introduction to Statistics: An Intuitive Guide For Analyzing Data and Unlocking Discoveries.* James D. Frost. ebook. p 256

[61] Wu K L 2012 Analysis of parameter selections for fuzzy c-means. *Pattern Recognit.* 45(1): 407–415

[62] Bagarello V, Castellini M, Di Prima S and Iovino M 2014 Soil hydraulic properties determined by infiltration experiments and different heights of water pouring. *Geoderma* 213: 492–501

[63] Lim D K and Kolay P K 2009 Predicting hydraulic conductivity (k) of tropical soils by using artificial neural network (ANN). *J. Civ. Eng. Sci. Technol.* 1(1): 1–6

[64] Arshad R R, Sayyad G, Mosaddeghi M and Gharabaghi B 2013 Predicting saturated hydraulic conductivity by artificial intelligence and regression models. *ISRN Soil Sci.* 1–8