Computers and Electronics in Agriculture 141 (2017) 327–339

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Using soil easily measured parameters for estimating soil water capacity: Soft computing approaches

Jalal Shiri [a], Ali Keshavarzi [b,*], Ozgur Kisi [c], Sepideh Karimi [a]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Laboratory of Remote Sensing and GIS, Department of Soil Science, University of Tehran, P.O. Box: 4111, Karaj 31587-77871, Iran
[c] School of Natural Sciences and Engineering, Ilia State University, Tbilisi, Georgia

ARTICLE INFO

Article history:
Received 3 June 2017
Received in revised form 8 August 2017
Accepted 9 August 2017
Available online 16 August 2017

Keywords:
Soft computing models
K-fold testing
Soil parameters
Soil moisture

ABSTRACT

The current study examines the applicability of six different soft computing approaches, gene expression programming (GEP), neuro-fuzzy (NF), support vector machine (SVM), multivariate adaptive regression spline (MARS), random forest (RF), and model tree (MT) techniques in modeling two important soil water capacity parameters, field capacity (FC) and permanent wilting point (PWP). Geometric mean particle-size diameter (dg), soil bulk density (BD), clay and silt obtained from 192 soil samples were introduced as input variables to the applied techniques and k-fold testing procedure was used for better comparison of the soft computing models. The best accuracy was provided by the NF models followed by the GEP, while the MT approach gave the worst estimates. The performances accuracies of the soft computing models in estimation of PWP parameter were higher than those in the FC estimation. Further, the soft computing approaches were compared with the traditional multi-variable linear regression (MLR) as well as the previously developed pedotransfer functions (PTFs) and the better FC and PWP estimates which confirms the superiority of the soft computing approaches. The NF model increased the performance of the best PTF (Aina-Periaswamy) by 33% with respect to *GMER* in FC estimation while the *SI* statistics of the best PTF (Ghorbani-Homaee) was decreased by 50% using the soft computing model. The performance of the best PTF (Aina-Periaswamy) with respect to *GMER* was increased by 74% in PWP estimation while the *SI* statistics of the best PTF (Dijkerman) was decreased by 99% using the soft computing model.

1. Introduction

Information on soil hydraulic properties, e.g. soil water content is very necessary in water and solute transport as well as heat and mass transfer in soils (Cornelis et al., 2001). Accurate knowledge about the soil available water capacity (AWC) is very important in various environmental issues including irrigation scheduling, land drainage and reclamation, analyzing soil biologic activity, surface runoff simulation, determining leaching requirement/fraction and crop growth simulation as well as different biophysical models (Rab et al., 2011). Since soil serves as a water circulator, the precise information on its moisture content will be very necessary for better managing the fertilizers application so that no excess runoff of these materials (which would be of high risk for the surface water environments) can be produced.

The AWC is defined as the difference between field capacity (FC) and permanent wilting point (PWP) (Waller and Yitayew, 2016). FC is the amount of soil moisture content held by the soil after the gravitational water was drained from the soil. It is indeed the bulk moisture content retained in the soil at −0.33 bar of hydraulic head (Veihmeyer and Hendrickson, 1931). PWP is defined as a minimum moisture content of a soil which is needed for the crop survival and if the water content decreases lower than PWP, a plant wilts and can no longer recover itself (Veihmeyer and Hendrickson, 1928). In-situe measurement of the FC and PWP moisture contents is very costly and time consuming, so numerous investigations have tried to relate these points to soil easily measured variables (Botula et al., 2012). A survey of the literature shows that soil easily measured variables, e.g. geometric mean particle-size diameter (dg), soil bulk density (BD), geometric standard deviation of soil particles (σ_g) and soil separates (clay, silt, sand) have been used to estimate soil FC and PWP (e.g. Aina and Periaswamy, 1985; Dijkerman 1988; Rab et al., 2011; Mohanty et al., 2015). Meanwhile, the soft computing approaches have been also applied for mapping the

* Corresponding author.
E-mail address: alikeshavarzi@ut.ac.ir (A. Keshavarzi).

http://dx.doi.org/10.1016/j.compag.2017.08.012
0168-1699/

input-output relationships between the FC, PWP and the soil easily measured variables.

Borgesen and Schaap (2005) applied neural networks (NN) to estimate soil water content at different pressures and found that introducing soil organic matter and BD as input vectors improves the modeling accuracy. Merdun et al. (2006) applied NN and multi-variable linear regression (MLR) techniques for estimating soil FC and PWP using 195 soil samples and found that the soil BD and dg are the most influential parameters on FC and PWP. Ahmad et al. (2010) utilized remote sensing data for estimating soil moisture through support vector machine (SVM) technique and found that SVM model performs better than NN and MLR models. Ostovari et al. (2015a) applied Mamdani fuzzy inference system and regression tree techniques for estimating FC using 210 soil samples and introduced the soil clay content, BD and dg as input parameters. The obtained results showed the regression tree's superiority to the fuzzy system. Ostovari et al. (2015b) applied MLR technique to relate the soil FC and PWP to the soil easily measured variables using 255 soil samples and confirmed the superiority of their developed regression-based relations to the other published relations. Based on their results, FC and PWP are mainly affected by the clay and dg. The literature review by the authors showed that there are only limited applications of the soft computing models for modeling soil FC and PWP. Nevertheless, most of the existing literatures have applied a single data set assignment, where the developed models have been trained using a part of the available data and tested using the rest of the available patterns, which might lead to partially valid results (Shiri et al., 2014a, 2014b). The present paper will focus on application of the gene expression programming (GEP), neuro-fuzzy (NF), SVM, multivariate adaptive regression spline (MARS), random forest (RF), and model tree (MT) techniques for estimating these points. Further, a multi-variable linear regression model will be applied and compared with the soft computing techniques. The most robust k-fold testing data assessing scenario will be applied for training and testing the models, where all the available input-target patterns are involved in both the training and testing stages, so there would be no "unseen" part of the data (Roushangar et al.,2014; Shiri et al., 2015, 2017).

2. Materials and methods

2.1. Study area and used data

Data from Mohr plain, Fars province, located in Southwest Iran [between the latitudes of 27°25′N to 27°59′N and longitudes of 52°21′E to 53°05′E with an area about 1900 km^2] were utilized in the current paper for establishing and evaluating the applied models. Fig. 1 shows the geographical position of the studied area. The main land uses are pastures and irrigated farming across the Mehran River.

After preliminary studies of topographic maps (1:25,000), study location was appointed. A simple random sampling scheme was designed using ArcGIS 10.2.2 software for an appropriate determination of soil sampling areas to consider spatial variation of the parameters affecting the field capacity (FC) and permanent wilting point (PWP) in the study region.

A total of 250 soil samples were obtained from two-first vertical depths (0–30 and 30–60 cm depth) of 125 representative soil profiles. In order to investigate the relation between FC and PWP with easily measurable properties and complete the objectives of this study, out of 192 soil samples from two-first vertical depths were selected randomly to design this research.

Depths were assigned to a soil textural class determined by the fractions of each soil separates (sand, silt, and clay) presence in a soil as indicated by the USDA textural triangle (Schoeneberger et al., 2002).

The sampling sites were designed to cover equally the entire area and to incorporate different soil and land use types. The collected disturbed soil samples were air dried, crushed and sieved using 2 mm sieve size. Large plant material and pebbles were separated and discarded.

Rates of clay (<0.002 mm), silt (0.002–0.05 mm), and sand (0.05–2 mm) particles were measured via sieving and sedimentation technique (Gee and Bauder, 1986). The clod method (Blake and Hartge, 1986) was utilized for determining bulk density (BD) with triple replications. The moisture contents at field capacity and wilting point were determined with a pressure plate apparatus at −33 and −1500 kPa, respectively (Cassel and Nielsen, 1986). Water saturation percentage and calcium carbonate equivalent (CCE) were determined using standard methods (Sparks et al., 1996). The dg (mm) and σ_g were calculated based on three particle size fractions (clay, silt, and sand content) as (Shirazi and Boersma, 1984):

$$d_g = \exp\{0.01[P_{sand} \cdot \mathrm{Ln}(d_{sand}) + P_{silt} \cdot \mathrm{Ln}(d_{silt}) + P_{clay} \cdot \mathrm{Ln}(d_{clay})]\} \quad (1)$$

Table 1 summarizes the statistical parameters of the used data set. From the table, it can be seen that the PWP has negative high skewed distribution (Skewness = −1.279). Differences between the maximum and minimum values are high for the FC and PWP (37.030% for FC and 18.064% for PWP). The geometric mean particle-size diameter (dg) presents the highest variability in terms of the coefficient of variations, skewness and kurtosis (1.752, 5.04, and 30.15, respectively). Among the soil separates, silt and sand present the maximum and minimum variations, respectively.

Table 2 sums up some previously published pedotransfer functions of FC and PWP estimation. As can be seen from these functions, soil separates (clay, sand and silt), BD and dg have been generally used for estimating FC and PWP. Alike to these functions and based on statistical analysis of the available data (not presented here), it was found that the soil clay, silt, BD and dg are the most influential parameters on FC and PWP, so they were utilized as input parameters of the applied soft computing models.

2.2. Data splitting and model assessment

A k-fold testing data assignment procedure was adopted here to feed the applied models with the input-target matrixes, so the complete data was divided into 10 subsets and the models were trained and tested each time using a portion of available patterns. Using this procedure, all the available input-target patterns were seen by the models for constructing the final estimation model. Accordingly, the GEP, NF, SVM, MARS, RF and MT models were trained and tested 60 times (6 techniques * 10 folds). Assessing the models' performance accuracy through k-fold testing would avoid getting partially valid conclusions which might be drawn down using traditional data management scenarios (Marti et al., 2013; Shiri et al., 2014a) as no any unseen input patterns would be remained in models' development.

Assessing the performance accuracy of the employed models was carried out using the geometric mean error (*GMER*), and the scatter index (*SI*) statistical criteria:

$$GMER = \mathrm{Exp}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{Ln}\left(\frac{\theta_{im}}{\theta_{io}}\right)\right] \quad (2)$$

$$SI = \frac{RMSE}{\overline{\theta_o}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_{io}-\theta_{im})^2}}{\overline{\theta_o}} \quad (3)$$

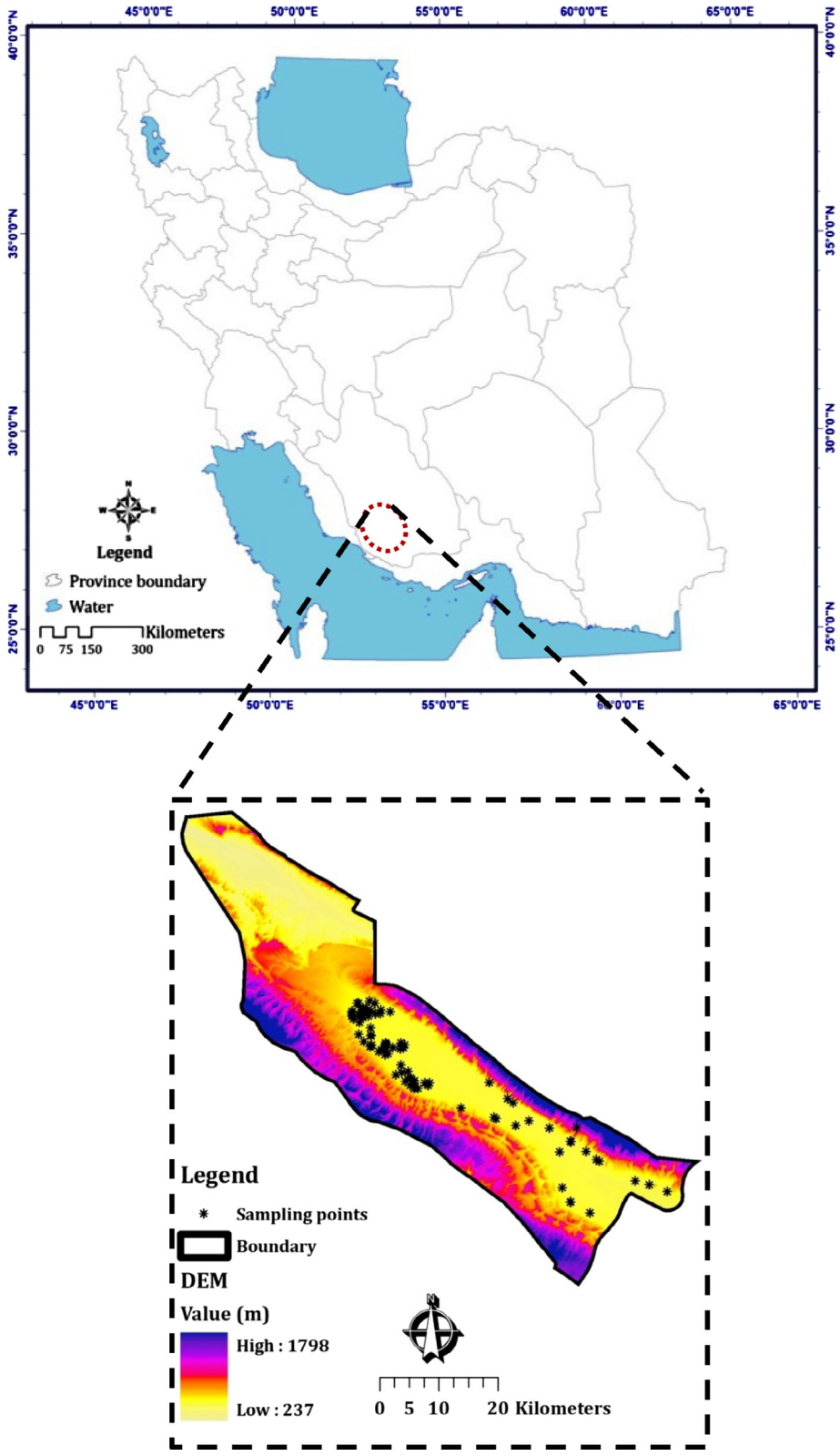


Fig. 1. Location of study area showing sampling points and position of Mohr region, Iran.

In these equations, θ_o represents the observed moisture (FC or PWP) value at the *i*th observation step, θ_m shows the corresponding estimated θ value, n denotes the number of patterns, $\overline{\theta_o}$ is the mean observed moisture value and *RMSE* stands for the root mean square error. The weighted *RMSE* (called as *SI*) is a dimensionless indicator which can give a fair comparison among the models when they rely on different targets (e.g. soil characteristics taken from different locations). This is very decisive as the effect of the targets' magnitude is ignored taken into consideration their mean values in *SI* calculation. Nonetheless, the *GMER* indicator presents the ratio

Table 1
Statistical parameters of the used data set.

	Clay (%)	Silt (%)	Sand (%)	CCE (%)	BD (g/cm^3)	PD (g/cm^3)	dg (mm)	σ_g (–)	FC (%)	PWP (%)
Maximum	65.440	44.000	92.720	48.700	1.964	2.780	0.682	23.391	45.605	22.084
Minimum	4.280	3.000	6.560	6.140	1.134	2.190	0.004	4.551	8.575	4.020
Mean	39.045	25.753	35.192	17.660	1.408	2.567	0.048	17.095	28.104	16.573
Standard deviation	11.868	7.025	16.394	6.387	0.131	0.084	0.084	3.733	7.248	3.639
Coefficient of variation	0.304	0.273	0.466	0.362	0.093	0.033	1.752	0.218	0.258	0.220
Skewness	−0.500	−0.215	0.877	1.316	0.446	−0.226	5.044	−0.821	0.010	−1.279
Kurtosis	0.071	0.502	0.927	2.982	0.906	1.723	30.152	0.522	−0.313	1.443

Note – CCE: Calcium Carbonate equivalent; BD: Bulk density; PD: Particle density; dg: Geometric mean particle-size diameter; σ_g: Geometric standard deviation of soil particles; FC: Field capacity; PWP: Permanent wilting point.

Table 2
Summary of the previously published pedotransfer functions for FC and PWP estimation.

Model	Expression	Reference
Aina-Periaswamy	$FC = 0.6788 - 0.0055 * sand - 0.0013 * BD$ $PWP = 0.00213 + 0.0031 * clay$	Aina and Periaswamy (1985)
Dijkerman	$FC = 0.3697 * BD + 0.0035 * sand * BD$ $PWP = 0.007 * BD + 0.0039 * clay * BD$	Dijkerman (1988)
Oliveira	$FC = 0.00333 * BD * silt + 0.00387 * BD * clay$ $PWP = 0.00038 * sand + 0.00153 * silt + 0.00341 * clay - 0.030861$	Oliveira et al. (2002)
Ghorbani-Homaee	$FC = 0.01 * (15.6 - 0.323 * sand + 16.9 * BD)$ $PWP = 0.01 * (6.627 + 0.315 * clay)$	Ghorbani-Dashtaki and Homaee (2004)
Ghanbarian-Millan	$FC = 0.401 - 0.165 * dg - 0.079 * BD + 0.003 * clay$ $PWP = 0.17 - 0.072 * dg - 0.052 * BD + 0.004 * clay$	Ghanbarian and Millan (2010)

between the target-output values which provides good insight about models central tendency. The applied indices were determined for the complete series for patterns (pooling the simulations of each data set) as well as split up for each test stage (one-tenth of the patterns matrix).

2.3. Gene expression programming (GEP)

Using a "parse tree" architecture, Genetic Programming (GP) (Koza, 1992), searches its solutions and is able to extract explicit expressions between the input-target set of a specified problem by using various operators. Alike to GP, gene expression programming (GEP) chooses the best controlling expressions using fitness magnitudes and brings genetic variation by utilizing a single or multiple genetic operators (Ferreira, 2006). One of the superiorities of GP (i.e. GEP) is in providing explicit formulation of the input-target relations in the studied problem. More detailed information regarding modeling procedure of GEP might be read in e.g. Ferreira (2006).

GeneXpro program (Ferreira, 2006) was used here for GEP runs, the procedure of soil *FC* and *PWP* modeling with which is as follows (Ferreira, 2006).

(i) Selection of the fitness function: Numerous absolute- and relative-error based fitness functions were evaluated in modeling soil FC and PWP as listed in Table 3. The *SI* values given in the table demonstrated that the root relative squared error (*RRSE*) was the best fitness function among

Table 3
Evaluating different fitness functions using the quadruple-input models.

Seletion of best fitness function

Fitness function based on the absolute error	*FC* *SI*	*PWP* *SI*	Fitness function based on the relative error	*FC* *SI*	*PWP* *SI*
Absolute error with selection range*	0.197	0.155	Relative error with selection range*	0.212	0.133
Absolute/hits	0.218	0.178	Relative/hits	0.231	0.189
Mean squared error (MSE)	0.210	0.137	r-MSE*	0.226	0.177
Root mean squared error (RMSE)	0.110	0.132	r-RMSE	0.216	0.148
Mean absolute error (MAE)	0.118	0.107	r-MAE	0.221	0.152
Relative squared error (RSE)	0.129	0.105	r-RSE	0.229	0.160
Root relative squared error (RRSE)	0.104	0.077	r-RRSE	0.211	0.129
Relative absolute error (RAE)	0.132	0.115	r-RAE	0.230	0.161

Selection of best operation function set

Function	Definition	*SI* (for *FC*)	*SI* (for *PWP*)
F1	$\{+, -, \times, \div\}$	0.199	0.123
F2	$\{+, -, \times, \div, \sqrt{}, x^2\}$	0.197	0.121
F3	$\{+, -, \times, \div, \sqrt{}, Power, Lnx, Logx, e^x, 10^x\}$	0.104	0.077
F4	$\{+, -, \times, \div, \sqrt[3]{}, \sqrt{}, \ln, e^x, x^2, x^3, \sin x, \cos x, Arctgx\}$	0.195	0.120

* Selection range was selected as maximum fitness; r: denotes relative error-based functions.

others. In some related literature (Kisi et al. 2012; Kisi and Shiri, 2012; Kisi et al., 2013), the *RRSE* has been also introduced as the optimum fitness function to be used in GEP modeling.

(ii) Selection of the input vectors and function sets. The input vectors (GEP terminals) in the current study for modeling FC and PWP are clay, silt, BD and dg. Several function sets listed in Table 3 were examined to choose the best one among them. Based on the *SI* values the function set F3 with addition linking function was ranked as the best set among other evaluated sets.

(iii) Defining chromosomal architecture: different values of head length and genes per chromosomes were evaluated (not presented here), from which the head size = 8 and number of genes = 3 found to be the optimum choices.

(iv) Genetic operators: GeneXpro default operators were used here as advised by literature (Number of chromosomes: 30, mutation rate: 0.044, inversion rate: 0.1, one point recombination rate: 0.3, two point recombination rate: 0.3, gene recombination rate: 0.1, gene transposition rate: 0.1, insertion sequence transposition rate: 0.1, root insertion sequence transposition: 0.1) (e.g. Shiri and Kisi, 2011).

2.4. Neuro-fuzzy system (NF)

NF is a combination of an adaptive neural network and a fuzzy inference system (FIS). The parameters of the FIS are identified using the NNs training algorithms. NF is based on the FIS, a crucial aspect is that the system must be interpretable in terms of fuzzy IF-THEN rules (Jang, 1993). NF recognizes a set of parameters via a hybrid learning procedure amalgamating back propagation and a least squared error approach. The neuro-fuzzy method applied in the present research executes the Sugeno's fuzzy approach (Takagi and Sugeno, 1985) for deriving the values for the target parameter from input parameters.

The usual grid partitioning identification method was employed in the present paper. One of the important steps by using this method involves selecting the appropriate membership function (MF) and its numbers (Kisi et al., 2012; Kisi and Shiri, 2012). Based on a trial and error procedure, 2, 3 or 4 numbers of MFs were found to be appropriate in the developed NF models. In Table 4, a comparison between different MFs has been presented, which demonstrated that using the triangular membership function gives the most accurate results among the other MFs, confirming the outcomes of Russel and Campbell (1996), where they have stated that this type of membership functions is commonly applied one in practical issues. However, this is a case sensitive object where a specified NF model is used to simulate a certain problem, so different MFs might have different results when applied to various problems, as reported in the literature (Vernieuw et al., 2005). The output MF was also chosen as "linear" since it surpassed the constant MF.

2.5. Support vector machine (SVM)

Support vector machine (SVM) is one of the novel soft computing algorithms which have been widely applied in environmental issues. It majorly finds lots of utilizations in forecasting, pattern recognition, regression analysis and classification. SVM was firstly developed by Vapnik et al. (1997), and is based on statistical machine learning procedure and structural risk minimization. It minimizes the upper bound generalization error instead of local training error, providing SVM a greater capability to generalize, which is the aim of statistical learning (Vapnik et al., 1997; Gunn, 1998).

The regression-SVM type1 was employed in the current study because its superiority to other types was demonstrated in previous studies (Shiri et al., 2014b). The difference between two types of regression-SVM is corresponded to their error functions (Vapnik et al., 1997). Using a trial and error procedure, SVM constants were selected as 8 (capacity) and 0.14 (epsilon). Linear, sigmoid, polynomial and radial basis kernel functions were used and evaluated and the radial basis kernel (Gamma = 0.25 and 0.11 for FC and PWP modeling, respectively) presented the best results (Table 4). Maximum number of iterations was found as 1200 (iteratively) and the models were stopped at error magnitudes of 0.004.

2.6. Multi-variate adaptive regression spline (MARS)

MARS is a non-parametric regression method which can be regarded as an accompaniment of linear models which automatically simulates the nonlinearities and interactions between

Table 4
SI values of sample evaluated NF, MARS, and SVM models.

	Structure		*SI*	
	FC	PWP	FC	PWP
NF models				
NF1 (Triangular MFs)	3, 4, 3, 4	3, 4, 3, 4	0.120	0.043
NF2 (Trapezoidal MFs)	3, 2, 3, 3	2, 2, 2, 2	0.142	0.098
NF3 (Generalized Bell MFs)	4, 3, 2, 3	2, 3, 2, 4	0.143	0.095
NF4 (Gaussian MFs)	2, 3, 3, 2	4, 2, 3, 2	0.123	0.050
NF5 (Two Gaussian MFs)	3, 3, 3, 3	2, 3, 2, 3	0.133	0.059
MARS models				
MARS1	20, 2, 2, 0.0005	20, 2, 2, 0.0005	0.172	0.129
MARS2	19, 3, 2, 0.0004	19, 3, 2, 0.0004	0.125	0.108
MARS3	18, 2, 2, 0.0002	18, 2, 2, 0.0002	0.172	0.128
MARS4	18, 4, 2, 0.0002	18, 4, 2, 0.0002	0.125	0.110
MARS5	25, 4, 2, 0.0002	25, 4, 2, 0.0002	0.124	0.104
SVM models				
SVM1 (linear kernel)	–	–	0.197	0.128
SVM2 (polynomial degree3 kernel)	0.30	0.28	0.196	0.125
SVM3 (sigmoid kernel)	0.20	0.20	0.198	0.130
SVM4 (radial basis function kernel)	0.25	0.11	0.129	0.106

Note: structures of the models in NF and SVM, denote the number of membership functions, and control parameters (Gamma), respectively. Structures of MARS models represent maximum number of basis functions, degree of interactions, penalty, and threshold, respectively.

parameters (Friedman, 1991). Its algorithm includes a forward and backward stepwise plan. In the forward stepwise plan, the intemperate forward stepwise choose plan would make a complex and over-trained model after a number of splits (Andres et al. 2010), which will have a lower performance accuracy. So, the backward stepwise plan eliminates the nonobligatory variables among the previously chosen set. This function would project variable X to a new variable Y via utilizing the following two basic functions, utilizing a knot or value of a variable which defines a conjunction point along the range of inputs (Sharda et al. 2006; Adamowski et al., 2012). MARS utilizes piecewise linear basis functions with the form of $(x-t)_+$ and $(t-x)_+$. The suffix "+" stands for the positive part only. So:

$$(x-t)_+ = \begin{cases} x-t, & if\ x > t \\ 0, & otherwise \end{cases} \quad (4)$$

and

$$(t-x)_+ = \begin{cases} t-x, & if\ x < t \\ 0, & otherwise \end{cases} \quad (5)$$

General form of the MARS equation reads:

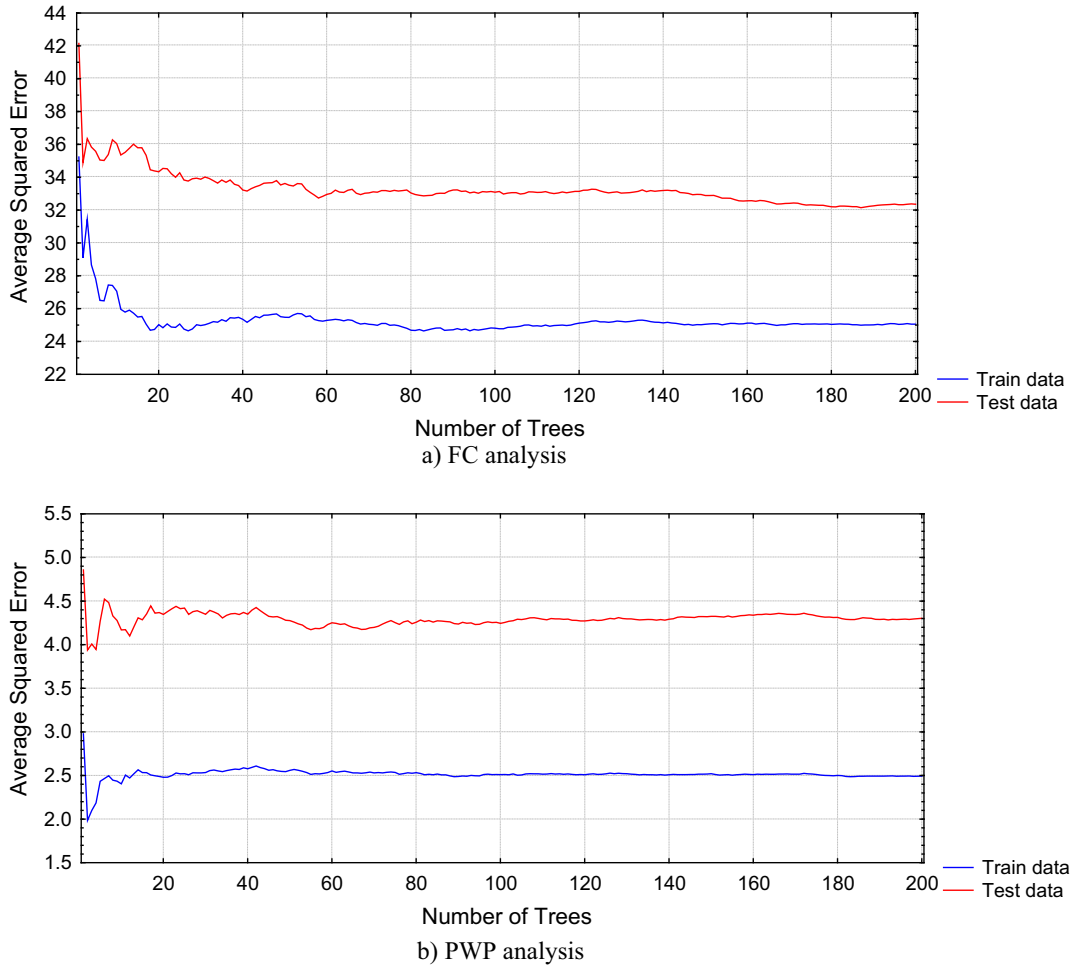


Fig. 2. Error variations vs. tree numbers of the RF models.

Table 5
Global error statistics of the applied models.

	GEP	NF	SVM	MARS	RF	MT	MLR	Aina-Periaswamy	Dijkerman	Oliveira	Ghorbani-Homaee	Ghanbarian-Millan
FC												
SI	0.104	0.100	0.129	0.124	0.138	0.180	0.209	0.205	0.349	0.220	0.201	0.225
R^2	0.838	0.847	0.751	0.766	0.714	0.508	0.414	0.472	0.434	0.359	0.347	0.455
GMER	0.994	0.995	0.990	0.990	0.979	0.964	0.786	0.747	0.436	0.164	0.014	0.314
PWP												
SI	0.077	0.043	0.106	0.104	0.105	0.166	0.230	4.757	2.818	2.894	3.315	2.969
R^2	0.873	0.960	0.764	0.768	0.784	0.531	0.411	0.312	0.376	0.336	0.312	0.270
GMER	0.997	0.999	0.990	0.990	0.989	0.969	0.738	−0.573	−0.222	−0.454	−0.549	−0.106

Table 6
t-Test results of the applied models.

	FC		PWP	
	t-Statistic	Resultant significance level	*t*-Statistic	Resultant significance level
GEP	0.194	0.888	0.002	0.990
NF	−0.076	0.939	0.0003	0.999
SVM	0.705	0.488	0.760	0.447
MARS	0.713	0.476	0.758	0.460
RF	0.733	0.424	0.798	0.400
MT	0.780	0.400	0.887	0.324
MLR	0.787	0.398	0.892	0.320

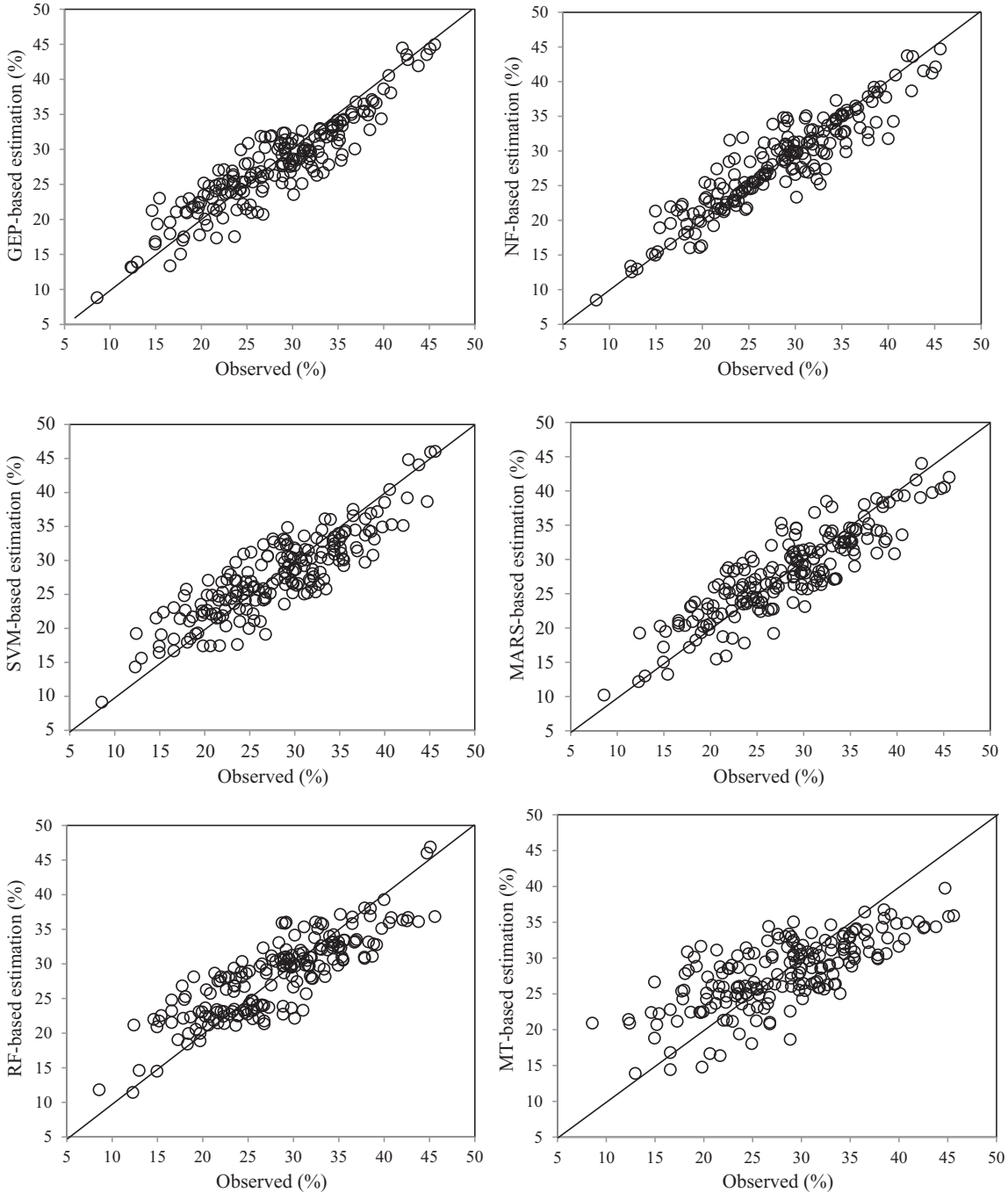


Fig. 3. Observed vs. estimated values of FC using the soft computing models.

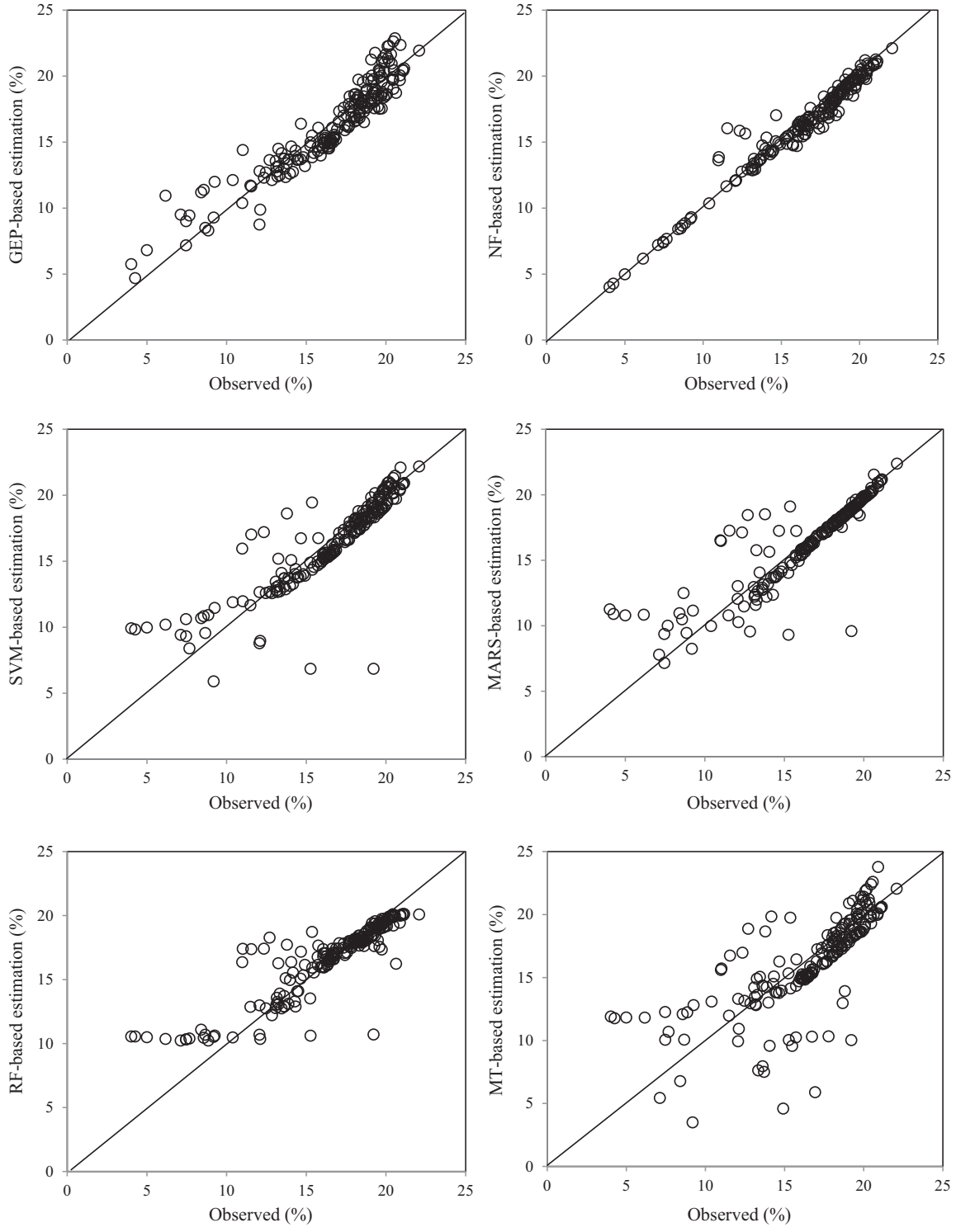


Fig. 4. Observed vs. estimated values of PWP using the soft computing models.

$$\widehat{f}(x) = \sum_{i}^{k} c_i B_i(x) \tag{6}$$

Which is indeed a weighted some of function $B_i(x)$. c_i shows the constant coefficients determined through minimizing the residual sum of squares (standard linear regression). The coefficients might be assumed as weights which depict the importance of each variable. The optimum MARS parameters were determined by a trial and error method. Table 4 presents the *SI* values of the sample evaluated MARS architectures. Among the evaluated MARS models for estimating FC and PWP, MARS5 gave the lowest error magnitudes.

2.7. Random Forest (RF)

Random Forests (RF) is a group learning algorithm which manages high-dimension regression problems. It is a tree-based group approach, where all trees are dependent of a collection of random

variables, and the forest is grown from many regression trees put together and from a group (Breiman, 2001). The final decision is resulted via averaging the output, after fitting single trees in ensemble (bagging procedure). The bias of the bagged trees is the same as that of the single trees, while the variance is decreased by reduction in the correlation between trees (Hastie et al., 2009).

Various tree numbers were analyzed for choosing the best random forest method (Fig. 2). From the figure, increasing the tree numbers beyond 80 makes lower variations in the average squared error values for both the FC and PWP estimation cases. Nonetheless, 15 cycles were found as the optimum cycle number of the mean error calculation, via a trial-error process. In a same way, the percentage decrease in training error was observed as 5%, minimum child node size to stop (for controlling the smallest permissible number in a child node, for a split to be applied) as 5, and the maximum number of levels (the depth of the tree as measured from the root node) as 10.

2.8. Model Tree (MT)

MT derives the conceptual knowledge of the classification and regression trees. These techniques are applied for solving the problem by separating it into different sub-domains tasks. The results will be a combination of these sub-domains. The difference between the classification trees and the MT technique is in having a numeric value rather than a class label in connection with the leaves. MTs split the total input domain into sub-domains and a multivariable linear regression model is applied for each sub-domain (Quinlan, 1992; Wang and Witten, 1977). MT models can be utilized for solving problems of continuous class to get a structural representation of the data sets utilizing the piecewise linear models to simulate nonlinear relationships. Based on the domain-splitting criterion, several approaches such as M5 model has been frequently employed to generalize a MT technique (Quinlan, 1992; Wang and Witten, 1977). More details on MT can be read in e.g. Goyal and Ojha (2011), Goyal (2014) and Mesbahi et al. (2017).

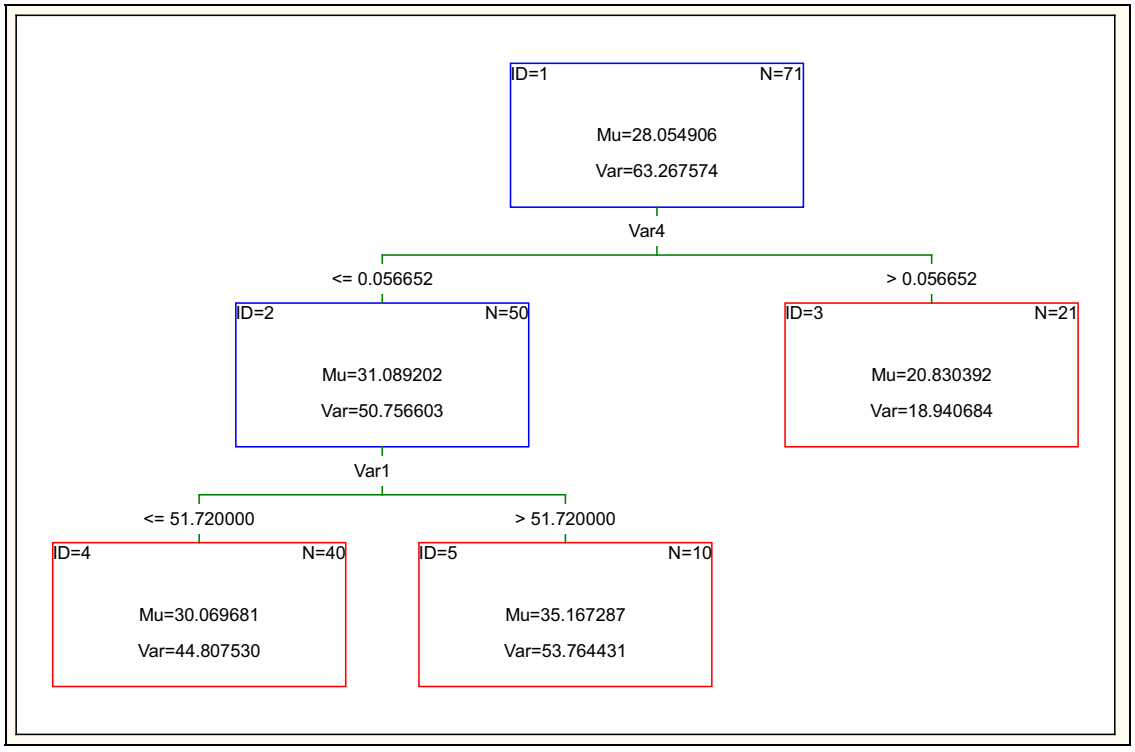
3. Results and discussions

Global error statistics of the GEP, NF, SVM, MARS, RF, MT and MLR approaches have been presented in Table 5 in terms of *SI*, R^2 and *GMER* indices. The presented values of these indices show the global performance accuracy of the employed approaches that have been determined by using the complete measured-estimated patterns. Analyzing the error statistics of the models in this table reveals that the NF model outperforms the other applied soft computing models with the highest R^2 (0.874 and 0.960 for FC and PWP, respectively) and *GMER* (0.995 and 0.999 for FC and PWP, respectively) and the lowest *SI* (0.10 and 0.043 for FC and PWP, respectively) magnitudes. GEP models are ranked as the second models for both FC and PWP estimations with some slight difference of performance accuracy with MARS and SVM models. In case of PWP modeling, RF gives promising results when its performance compared with those of the SVM and MARS, while in FC modeling, it comprises higher magnitudes of errors. The linear MT model (M5) showed poorer results then the rest, for both the modeling cases, which might be explained by its linear nature, where a linear relationship is assumed between the input-target attributes, while the existing relations between soil parameters are usually nonlinear. Finally, the traditional MLR model gave the worst results with the highest error magnitudes. Although very fine pores are involved in moisture suction at high matric potential values (e.g. PWP point) which can make its simulation difficult, the performance accuracy of the models in PWP estimation is better than those of the FC estimation. Similar outcomes were also observed by Ostovari et al. (2015b) where they proposed pedotransfer functions of FC and PWP estimations for some soils from Iran and reported higher R^2 values for PWP pedotransfer function.

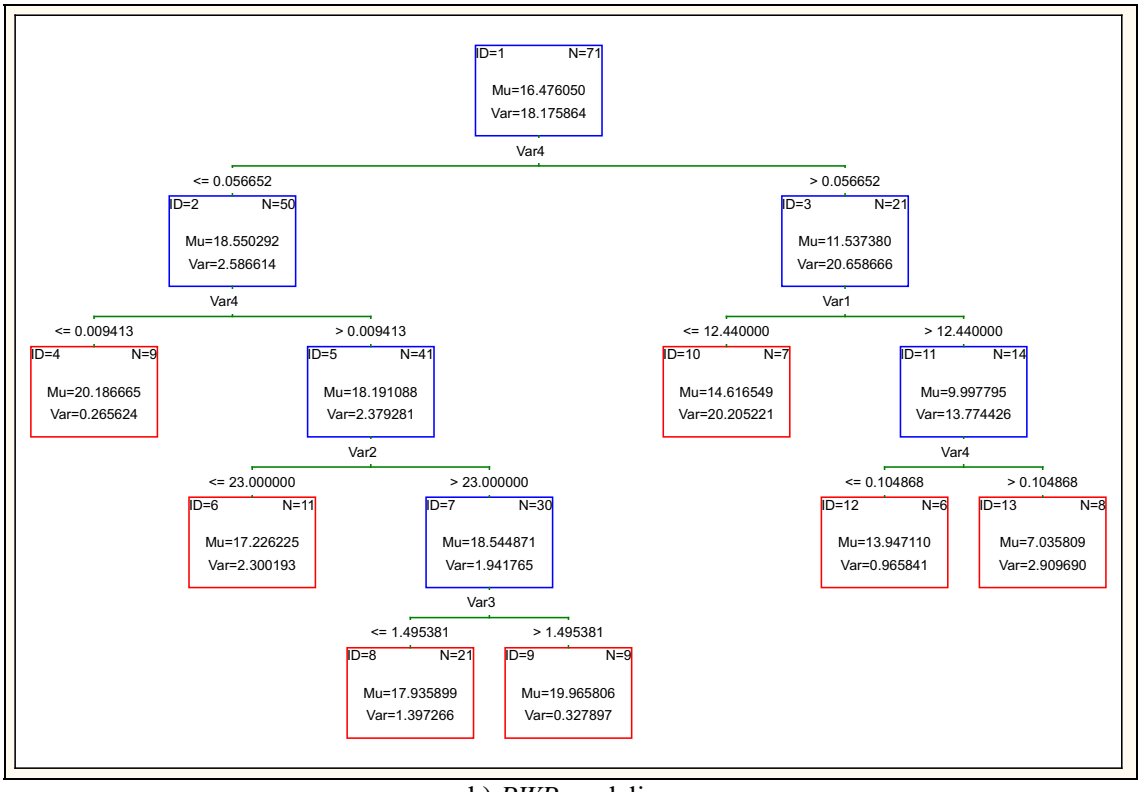
Comparison with previous studies indicates that all the applied soft computing approaches perform better than the MLR (developed in the present study) and the applied existing pedotransfer functions (PTFs) of Aina-Periaswamy, Dijkerman, Oliveira, Ghorbani-Homaee and Ghanbarian-Millan. Among the PTFs, Aina-Periaswamy showed the best accuracy with respect to *GMER* in estimating FC and PWP. According to the *SI* criterion, Ghorbani-Homaee and Dijkerman provided the best estimates of the FC and PWP, respectively. In FC estimation, the best soft computing NF approach increased the *GMER* accuracy of the best PTF (Aina-Periaswamy) by 33% while the *SI* accuracy of the best PTF (Ghorbani-Homaee) was increased by 50% using the NF model. For the PWP estimation, the accuracy of the best PTF (Aina-Periaswamy) was increased with respect to *GMER* by 274% while the *SI* accuracy of the best PTF (Dijkerman) was increased by 99% using the best NF model. This also indicates that the soft computing methods are more successful in estimating PWP when compared to FC. As can be seen from Table 2, the PTFs generally involve mostly linear equations and these cannot adequately model the nonlinear FC and PWM. PWP has more nonlinearity than the FC and therefore the difference between the soft computing and PTFs is higher for the PWP when compared to FC. It is relevant to note here that the applied existing pedotransfer functions have been developed using data from other regions, so the applied soft computing models here might present the crucial advantage of being trained using local patterns. Further, even though the developed MLR is relying on the local patterns applied in the present

Table 7
Mathematical expressions of the GEP, MARS, MT and MLR models.

FC models	
GEP	$FC = \sqrt{(4.4525dg + silt) + clay - Ln(dg^2)} + \frac{BD - clay + 2.634}{BD} + BD + clay + 16.177dg^2 + dg^{silt} - Ln(BD)$
MARS	*FC* = 2.53369818333513e+001 − 3.21804053831195e−001 * max(0; 3.20000000000000e+001-*silt*) + 4.38766157881459e−001 * max(0; *clay*-3.27200000000000e+001) + 2.90633211544715e+001 * max(0; *BD* − 1.43300000000000e+000)
MT	$FC = 0.3422 * clay + 0.159 * silt + 15.332 * BD - 11.0059$
MLR	$FC = 0.82 * clay + 0.199 * silt + 15.33 * BD + 9.58dg - 13.9$
PWP models	
GEP	$PWP = \sqrt{BD(silt + dg + clay - 0.2975)} + 0.099clay * 10^{\log(BD)} + 2dg + BD$
MARS	*PWP* = 1.34428872081241e+001 + 1.22438933317267e+001 * max(0; *dg*-7.05620321236376e−002) + 1.12545237621318e+002 * max(0; 7.05620321236376e−002-*dg*) + 1.34843210529738e+001 * max(0; *BD*-1.50976200000000e+000) − 9.55169493891780e+000 * max(0; 1.50976200000000e+000-*BD*) − 3.39761788785992e−001 * max(0; 3.00000000000000e+001-*clay*)
MT	$PWP = -0.0258 * clay + 0.0372 * silt + 2.7552 * BD + 1.5438 * dg + 5.521$
MLR	$PWP = 0.24 * clay + 0.139 * silt + 10.3 * BD + 5.77 * dg - 11.1$



a) *FC* modeling



b) *PWP* modeling

Fig. 5. tree graphs of the RF model (Variables 1–4 denotes clay, silt, BD and dg, respectively; Mu = mean of data; Var = data variance).

study, its performance accuracy is poorer than the soft computing models which can be attributed to the linear linkage of the input-output attributes with this technique. So, the performance analysis of the MLR and the existing PTFs will not be additionally discussed in the subsequent sections of the paper as they failed to produce promising results.

The overall *SI* difference between the performance accuracy of the models is very small in case of the FC estimation, where the

Table 8
Correlations between soil parameters and errors of the applied models.

FC						PWP					
	Clay	Silt	BD	dg	FC		Clay	Silt	BD	dg	PWP
Clay	1.000	0.470	−0.279	−0.718	0.559	Clay	1.000	0.470	−0.279	−0.718	0.711
Silt	0.470	1.000	−0.034	−0.605	0.409	Silt	0.470	1.000	−0.034	−0.605	0.543
BD	−0.279	−0.034	1.000	0.175	0.115	BD	−0.279	−0.034	1.000	0.175	0.166
dg	−0.718	−0.605	0.175	1.000	−0.406	dg	−0.718	−0.605	0.175	1.000	−0.527
FC	0.559	0.409	0.115	−0.406	1.000	PWP	0.711	0.543	0.166	−0.527	1.000
GEP	−0.099	−0.031	−0.078	0.026	−0.034	GEP	−0.241	−0.105	0.128	0.207	−0.275
NF	0.190	0.059	−0.099	−0.168	0.052	NF	0.077	0.101	−0.151	−0.083	−0.178
SVM	−0.017	0.008	−0.141	−0.032	0.021	SVM	−0.344	−0.252	0.057	0.456	−0.210
MARS	0.048	−0.027	−0.081	0.005	0.027	MARS	−0.341	−0.238	0.061	0.408	−0.403
RF	0.083	−0.079	−0.064	−0.027	−0.060	RF	−0.327	−0.194	0.098	0.342	−0.420
MT	0.048	0.139	−0.131	−0.034	−0.122	MT	−0.215	−0.189	−0.130	0.217	−0.310

ΔSI between the NF model and GEP, MARS, SVM, 0.004, 0.029, 0.024, 0.038, and 0.080. Such differences in case of PWP estimation are considerably higher (0.034, 0.063, 0.061, 0.062 and 0.123). Comparing the performance of the soft computing methods with the existing PTFs shows that the soft computing models outperform the functions for both FC and PWP modeling. A reason behind this might be the study origin, for which soils the PTFs have been developed and validated. The obtained results were also additionally applied to a *t*-test (at a significant level of 95%) and the results were presented in Table 6. From the statistics it is seen that the NF models give the most accurate results with the lowest *t*-statistics and the highest significance level followed by the GEP models. Here also the MLR has the worst test results with respect to *t*-test among the soft computing approaches.

Figs. 3 and 4 present the scatterplots of the observed vs. simulated FC and PWP values using all the available patterns. As mentioned, the models have been assessed through a k-fold testing, so the simulations presented here are the amalgamated matrix of the all estimated values obtained through each k-fold testing stage. In both the cases, NF-bases outcomes show the lowest scatters between the observed-simulated values. MT provides the most scattered estimates for the FC and PWP. Its linear structure prevents MT from catching extreme values of the highly nonlinear soil water capacity parameters. The estimations corresponded to the FC are monotonously scattered around the identity function (the straight line drawn in each plot), while there are large scatters between the observed and estimated values for lower magnitudes of the PWP. However, as the moisture magnitudes in PWP are relatively small than those of the FC point, such larger scatters have not made higher magnitudes for error statistics (Table 5).

Mathematical expressions of the GEP, MARS, MT and MLR models are presented in Table 7. The RF tree models are also given in Fig. 5. Analyzing the models structures showed that GEP gives the highest weight to *dg* in modeling FC, while *BD* and *dg* have the same weights in PWP simulation. MARS and MT models ignore dg in their modeling for FC simulation and gave similar weight to the remains variables. Nonetheless, MARS ignores applying silt for PWP simulation and considers the *BD* and *dg* as the most influential parameters on PWP. It should be noted that the MT equation is very similar to that of the Oliveira et al. (2002) for modeling FC so that both are linear equations and they include silt, *BD* and clay. RF model performs slightly different than the others, where gives the highest weight to clay and *dg* and ignores *BD* and silt in FC modeling, while all introduced variables are incorporated in PWP estimation. Such difference might be explained through the differences between the models basic algorithms and assumptions.

Table 8 sums up the correlations between the soil parameters and the models' errors. The highest FC correlation values are observed with clay (0.559) followed by the silt (0.409) and dg (−0.406). PWP shows a high correlation with clay (0.711) and moderate correlations with silt (0.543) and dg (−0.527). Analyzing the correlations between the models errors and soil parameters confirms the statement given in the previous section about the variables importance.

Split up statistics per test stage of the employed techniques has been illustrated in Fig. 6. Both *SI* and R^2 indices show considerable variations among the test stages for all employed models. In FC modeling, MT has the highest *SI* values in all cases while the RF has lower R^2 than the MT in some cases (cases 1, 9 and 10). NF generally has the lowest *SI* and the highest R^2 in modeling FC except the cases 3, 4, 8 and 10 where the GEP provides better accuracy than the NF. In PWP modeling, the MT has the highest *SI* and the lowest R^2 values except two cases, 6 and 8, where GEP and RF show slightly worse accuracy than the MT, respectively. In general, a better accuracy was obtained from NF with respect to *SI* and R^2 in comparison to the other methods.

4. Conclusions

The study investigated the estimation of soil water capacity parameters, FC and PWP using six different soft computing approaches, namely, GEP, NF, SVM, MARS, RF and MT. To better compare the performance of the applied models, k-fold testing data assessing scenario was employed in which all the available input-target patterns are included in both training and testing stages. The clay, silt, BD and dg obtained from 192 soil samples were used as inputs to the soft computing models to estimate the FC and PWP parameters. NF model generally found to be better than the other models in estimating soil water capacity parameters while the MT provided the worst accuracy. The RF and MT could not sufficiently model the FC and PWP in most of the k-fold cases. The soft computing models provided much better accuracy in modeling PWP in comparison to the FC parameter.

The soft computing methods were also compared with multivariable linear regression (MLR) and some existing PTFs and they performed better than the MLR and PTFs. The best NF model increased the *GMER* accuracy of the best PTF (Aina-Periaswamy) by 33% in FC estimation while the *SI* of the best PTF (Ghorbani-Homaee) was decreased by 50% using the NF model. The *GMER* of the best PTF (Aina-Periaswamy) was increased by 274% in PWP estimation while the *SI* of the best PTF (Dijkerman) was decreased by 99% using the best NF model. A higher difference (49%) was obtained between the PTFs and soft computing models in PWP estimation in comparison to FC (41%).

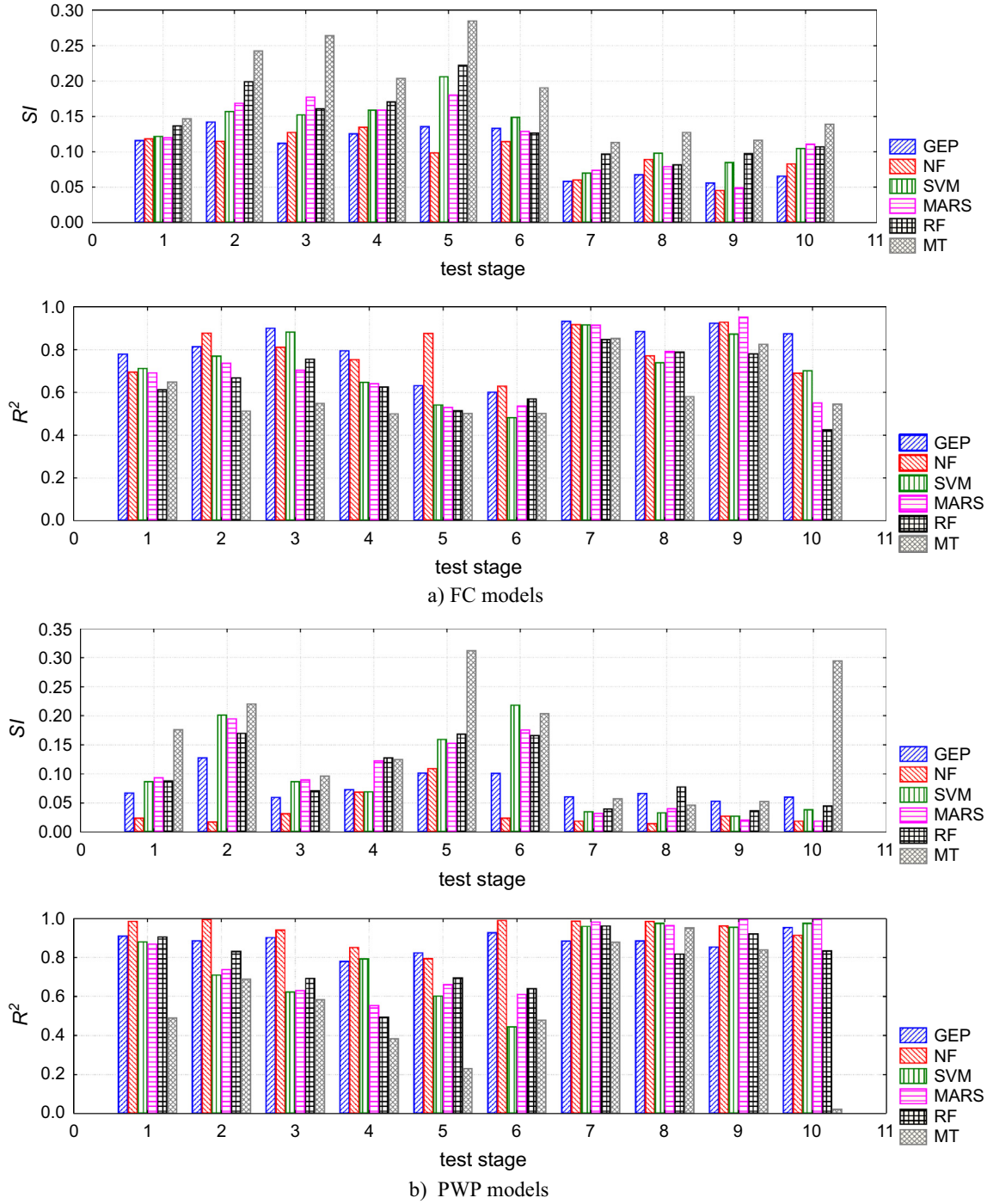


Fig. 6. Split per test stage statistics of the models.

Acknowledgement

This study was partially supported by Department of Soil Science, University of Tehran, Iran. The authors thank the editor and anonymous reviewers for their help in improving the quality of the manuscript.

References

Adamowski, J., Chan, H.F., Prasher, S.O., Sharda, V.N., 2012. Comparison of multivariate adaptive regression splines with coupled wavelet transform artificial neural networks for runoff forecasting in Himalayan micro-watersheds with limited data. J. Hydroinform. 14 (3), 731–744.

Ahmad, S., Kalra, A., Stephen, H., 2010. Estimating soil moisture using remote sensing data: a machine learning approach. Adv. Water Resour. 33, 69–80.

Aina, P.O., Periaswamy, S.P., 1985. Estimating available water-holding capacity of western Nigerian soils from soil texture and bulk density, using core and sieved samples. Soil Sci. 140, 55–58.

Andres, J.D., Lorca, P., de Cos Juez, F.J., Sánchez-Lasheras, F., 2010. Bankruptcy forecasting: a hybrid approach using fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). Expert Syst. Appl. 38, 1866–1875.

Blake, G.R., Hartge, K.H., 1986. Bulk density. In: Page, A.L. (Ed.), Methods of Soil Analysis, Part 1. American Society of Agronomy, Madison, Wisconsin.

Borgesen, C.D., Schaap, M.G., 2005. Point and parameter pedotransfer functions for water retention predictions for Danish soils. Geoderma 127, 154–167.

Botula, Y.D., Cornelis, W.M., Baert, G., Van Ranst, E., 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (D.R. Congo). Agric. Water Manag. 111, 1–10.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cassel, D.K., Nielsen, D.R., 1986. Field capacity and available water capacity. In: Klute, A. (Eds.), Methods of Soil Analysis. Part 1, 2nd ed. Agron Monogr 9, ASA and SSSA, Madison, WI, USA. pp. 901–926.
Cornelis, W.M., Ronsyn, J., van Meirvenne, M., Hartmann, R., 2001. Evaluation of pedotransfer functions for predicting the soil moisture retention curve. Soil Sci. Soc. Am. J. 65, 638–648.
Dijkerman, J.C., 1988. An ustult-aqualt-tropept catena in Sierra Leone, West Africa, II. Land qualities and land evaluation. Geoderma 42 (1), 29–49.
Ferreira, C., 2006. Gene Expression Programming: Mathematical Modeling by An Artificial Intelligence. Springer, Berlin, Heidelberg, New York, p. 478.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19, 1.
Gee, G.W., Bauder, J.W., 1986. Particle size analysis. In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 383–411.
Ghanbarian, A.B., Millan, H., 2010. Point pedotransfer functions for estimating soil water retention curve. Int. Agrophys. 24 (3), 243–251.
Ghorbani-Dashtaki, S., Homaee, M., 2004. Using geometric mean particle diameter to derive point and continuous pedotransfer functions. Eurasian Soil Sci. 30 (1–10), 10.
Goyal, M.K., 2014. Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression. Water Resour. Manage 28 (7), 1991–2003.
Goyal, M.K., Ojha, C.S.P., 2011. Estimation of scour downstream of a ski-jump bucket using support vector and M5 model tree. Water Resour. Manage 25 (9), 2177–2195.
Gunn, S.R., 1998. Support vector machines for classification and regression Technical Report. University of Southampton, England.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer, New York.
Jang, J.S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. 23 (3), 665–685.
Kisi, O., Dailr, A.H., Cimen, M., Shiri, J., 2012. Suspended sediment modeling using genetic programming and soft computing techniques. J. Hydrol. 450–451, 48–58.
Kisi, O., Shiri, J., 2012. River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques. Comput. Geosci. 43, 73–82.
Kisi, O., Shiri, J., Tombul, M., 2013. Modeling rainfall-runoff process using soft computing techniques. Comput. Geosci. 51, 108–117.
Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press, Cambridge, MA, p. 840.
Marti, P., Shiri, J., Duran-Ros, M., Arbat, G., Cartagena, F.R., Puig-Bargues, J., 2013. Artificial neural networks vs. gene expressions programming for estimating outlet dissolved oxygen in micro irrigation sand filters fed with effluents. Comput. Electron. Agric. 99, 176–185.
Merdun, H., Cinar, O., Meral, R., Apan, M., 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil Till. Res. 90, 108–116.
Mesbahi, M., Talebbeydokhti, N., Hosseini, S.-A., Afzali, S.H., 2017. External validation criteria and uncertainty analysis of maximum scour depth at downstream of stilling basins based on EPR and MT approaches. Iran. J. Sci. Technol. Trans. Civ. Eng. 41 (1), 87–99.
Mohanty, M., Sinha, N.K., Painuli, D.K., Bandyopadhyay, K.K., Hati, K.M., Reddy, K.S., Chaudhary, R.S., 2015. Modelling soil water contents at field capacity and permanent wilting point using artificial neural network for Indian soils. National Acad. Sci. Lett. 38 (5), 373–377.
Oliveira, L.B., Ribeiro, M.R., Jacomine, P.K.T., Rodrigues, J.J.V., Marques, F.A., 2002. Pedotransfer functions for the prediction of moisture retention and specific potentials in soils of pernambuco state (Brazil). Revista Brasileira de Ciência do Solo 26, 315–333.
Ostovari, Y., Asgarib, K., Cornelisc, W., Beigi-Harchegania, H., 2015a. Simple methods for estimating field capacity using Mamdani inference system and regression tree. Arch. Agron. Soil Sci. 61 (6), 851–864.
Ostovari, Y., Asgari, K., Cornelis, W., 2015b. Performance evaluation of pedotransfer functions to predict field capacity and permanent wilting point using UNSODA and HYPRES datasets. Arid Land Res. Manage. 29, 383–398.
Quinlan, J.R., 1992. Learning with continuous classes. In: Adams, Sterling (Ed.), Proceedings of AI'92. World Scientific, pp. 343–348.
Rab, M.A., Chandra, S., Fisher, P.D., Robinson, N.J., Kitching, M., Aumann, C.D., Imhof, M., 2011. Modelling and prediction of soil water contents at field capacity and permanent wilting point of dryland cropping soils. Soil Res. 49, 389–407.
Roushangar, K., Vojoudi, F., Shiri, J., 2014. Modeling river total bed material load discharge using artificial intelligence approaches (based on conceptual inputs). J. Hydrol. 514, 122–144.
Russel, S.O., Campbell, P.F., 1996. Reservoir operating rules with fuzzy programming. J. Water Resour. Plan. Manage. 122 (3), 165–170.
Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Broderson, W.D., 2002. Field Book for Describing and Sampling Soils, Version 2.0. NRCS-National Soil Survey Center, Lincoln, NE.
Sharda, V., Prasher, S.O., Patel, R.M., Ojavasi, P.R., Prakash, C., 2006. Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques. Agric. Water Manag. 83, 233–242.
Shirazi, M.A., Boersma, L., 1984. A unifying quantitative analysis of soil texture. Soil Sci. Soc. Am. J. 48, 142–147.
Shiri, J., Kisi, O., 2011. Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Comput. Geosci. 37 (10), 1692–1701.
Shiri, J., Marti, P., Singh, V.P., 2014a. Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning. Hydrol. Process. 28 (3), 1215–1225.
Shiri, J., Nazemi, A.H., Sadraddini, A.A., Landeras, G., Kisi, O., Fakheri Fard, A., Marti, P., 2014b. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput. Electron. Agric. 108, 230–241.
Shiri, J., Marti, P., Nazemi, A.H., Sadraddini, A.A., Kisi, O., Landeras, G., Fakheri Fard, A., 2015. Local vs. external training of neuro-fuzzy and neural networks models for estimating reference evapotranspiration assessed through k-fold testing. Hydrol. Res. 46 (1), 72–88.
Shiri, J., Keshavarzi, A., Kisi, O., Iturraran-Viveros, U., Bagherzadeh, A., Mousavi, R., Karimi, S., 2017. Modeling soil cation exchange capacity using soil parameters: assessing the heuristic models. Comput. Electron. Agric. 135, 242–251.
Sparks, D.L., Page, A.L., Helmke, P.A., Leoppert, R.H., Soltanpour, P.N., Tabatabai, M. A., Johnston, G.T., Summer, M.E., 1996. Methods of Soil Analysis. Soil Science Society of America, Madison, Wisconsin.
Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its application to modeling and control. IEEE Trans. Syst. Man Cybern. 15 (1), 116–132.
Vapnik, V., Golwich, S., Smola, A.J., 1997. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer, M., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems, 9, pp. 281–287.
Veihmeyer, F.J., Hendrickson, A.H., 1928. Soil moisture at permanent wilting of plants. Plant Physiol. 3 (3), 355–357.
Veihmeyer, F.J., Hendrickson, A.H., 1931. The moisture equivalent as a measure of the field capacity of soils. Soil Sci. 32 (3), 181–193.
Vernieuwe, H., Georgieva, O., De Baets, B., Pauwels, V.R.N., Verhoest, N.E.C., De Troch, F.P., 2005. Comparison of data-driven Takagi-Sugeno models of rainfall-discharge dynamics. J. Hydrol. 302 (1–4), 173–186.
Waller, P., Yitayew, M., 2016. Irrigation and Drainage Engineering. Springer International Publishing AG, Switzerland. 747 P.
Wang, Y., Witten, I.H., 1977. Induction of model trees for predicting continuous lasses. In: Proceedings of the Poster Papers of the European Conference on Machine Learning. Prague: University of Economics, Faculty of Informatics and Statistics.