

Rothamsted Repository Download

A - Papers appearing in refereed journals

Rastgou, M., Bayat, H., Mansoorizadeh, M. and Gregory, A. S. 2020.
Estimating the soil water retention curve - comparison of multiple
nonlinear regression approach and random forest data mining technique.
Computers and Electronics in Agriculture. 174, p. 105502.

The publisher's version can be accessed at:

- <https://dx.doi.org/10.1016/j.compag.2020.105502>

The output can be accessed at:

<https://repository.rothamsted.ac.uk/item/979z1/estimating-the-soil-water-retention-curve-comparison-of-multiple-nonlinear-regression-approach-and-random-forest-data-mining-technique>.

© Please contact library@rothamsted.ac.uk for copyright queries.

Manuscript Details

Manuscript number	COMPAG_2019_2240_R2
Title	Estimating the soil water retention curve: comparison of multiple nonlinear regression approach and random forest data mining technique
Article type	Research Paper

Abstract

This study evaluates the performance of the random forest (RF) method on the prediction of the soil water retention curve (SWRC) and compares its performance with those of nonlinear regression (NLR) and Rosetta-based pedotransfer functions (PTFs), which has not been reported so far. Fifteen RF and NLR-based PTFs were constructed using readily-available soil properties for 223 soil samples from Iran. The general performance of RF and NLR-based PTFs was quantified by the integral root mean square error (IRMSE), Akaike's information criterion (AIC) and coefficient of determination (R²). The results showed that the accuracy of the RF-based PTFs was significantly ($P < 0.05$) better than the NLR-based PTFs, and that the reliability of the NLR-based PTFs was significantly ($P < 0.01$) better than the RF-based PTFs and all of the Rosetta-based PTFs. The average values of the IRMSE, AIC and R² of the RF method were 0.041 cm³ cm⁻³, -16997.7, and 0.987, and 0.053 cm³ cm⁻³, -15547.5, and 0.981 for the training and testing steps of all PTFs, respectively, whereas the values for the NLR method were 0.046 cm³ cm⁻³, -16616.4, and 0.984, and 0.048 cm³ cm⁻³, -16355.6, and 0.983 for the training and testing steps, respectively. The PTF5 of the RF and NLR methods, with inputs of sand and clay contents, bulk density, and the water content at field capacity and permanent wilting point, had the greatest R² values (0.987 and 0.989, respectively), and the lowest IRMSE values (0.039 and 0.032 cm³ cm⁻³, respectively) compared to other PTFs for the testing step. Overall, the RF method had less reliability for the prediction of the SWRC compared to the NLR method due to overprediction, uncertainty of determination of forest scale and instability in the testing step. These findings could provide the scientific basis for further research on the RF method.

Keywords	pedotransfer functions; soil water retention curve; soil texture; soil structure; van Genuchten.
Corresponding Author	Hossein Bayat
Corresponding Author's Institution	Bu Ali Sina University
Order of Authors	Mostafa Rastgou, Hossein Bayat, Muharram Mansoorizadeh, Andrew Gregory
Suggested reviewers	Jorge Werneck Lima, Estela. N. Hepper, Sabit Ersahin, Masoud Davari, Hojat Emami

Submission Files Included in this PDF

File Name [File Type]

Cover letter.docx [Cover Letter]

Answers to the comments.docx [Response to Reviewers]

Revision changes marked-pic.docx [Review Reports]

Highlights.docx [Highlights]

Clean File.docx [Manuscript File]

conflict of interests.docx [Conflict of Interest]

Author statement.docx [Author Statement]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Dear Prof. He,

Thanks a lot for your helpful advice and the reviewers' useful comments and suggestions on our manuscript. We modified and revised the manuscript accordingly and details of the corrections are described in the “Answer to the comments” file, point by point. One of the co-authors is a native English speaker and he has thoroughly checked and corrected spelling and grammatical errors. Then two versions of the manuscript were resubmitted to your journal: a version of the revised manuscript showing the new/changed text using track changes and a clean version of the revised manuscript. It would be appreciated if you could please kindly let me know if there is any other deficiency with our manuscript. We look forward to your positive response.

With best regards

Hossein Bayat

Hossein Bayat: Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. Postal Address: Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran.

E-mail; h.bayat@basu.ac.ir. Tel: +98-918-8188378 and +98-81-34424189

1 Dear Prof. He,
2 Thanks a lot for your helpful advice and the reviewers' useful comments and suggestions on our
3 manuscript. We modified and revised the manuscript accordingly and details of the corrections
4 are described below point by point. One of the co-authors is a native English speaker and he has
5 thoroughly checked and corrected spelling and grammatical errors. Then two versions of the
6 manuscript were resubmitted to your journal: a version of the revised manuscript showing the
7 new/changed text using track changes and a clean version of the revised manuscript. It would be
8 appreciated if you could please kindly let me know if there is any other deficiency with our
9 manuscript. We look forward to your positive response.

10 With best regards

11 Hossein Bayat

12 **Note: Page numbers and line numbers that are given in this file are according to those of**
13 **the "Revision, changes marked" file.**

14

15

16

17

18

19

20

21

22

23

24 **Comments from the editors and reviewers:**

25 **-Reviewer 1**

26 -The authors replied to only partially to my comments. In fact, they replied to those comments
27 for Author but not to those for Editor. Did they not receive the comments for Editor? My
28 decision is still Major revision since many points raised before did not receive answers. I put
29 them again below. **Page and line numbers refer to the original version of the manuscript,**
30 **not the revised one.**

31 **Ans:**

32 We apologize for the inconvenience, but unfortunately we did not receive the comments in the
33 first round of Review. Now, we have modified and revised the manuscript according to your
34 comments and details of the corrections are described below point by point. The authors are
35 grateful for your comments in improving the content and structure of the manuscript.

36 First of all, the use of random forest to PTF is not completely new as may be deduced from the
37 manuscript (page 5, lines 88 and 89); in contrast, there are published papers that dealt with random
38 forest like Toth et al (2014), Araya et al (2019), Gunarathna et al (2019), and Szabo et al (2019).
39 Also, the authors gave few examples of the use of statistical and data mining techniques but only
40 one example for the nearest neighbor (page 4, line 75) as if it is the only published work whereas
41 there are many other examples like Botula et al (2013), Haghverdi et al (2015), Nguyen et al
42 (2017), Gunarathna et al (2019), etc.

43 **Ans:**

44 Thank you so much. A review of literatures (Toth et al. (2014), Araya et al. (2019), Gunarathna
45 et al. (2019), and Szabo et al. (2019)) revealed that the RF data mining technique has only been
46 applied to predict point-based PTFs of the SWRC including field capacity and permanent wilting

47 point or saturated hydraulic conductivity, but it has not been used for developing parametric-
48 based PTFs of the van Genuchten model parameters, so far. Finally, the review of literatures has
49 been modified completely as follows:

50 "So far, few studies have been carried out on the application of the RF method in soil science.
51 Tóth et al. (2014) applied the RF method to analyze the relationship between soil water content
52 at four matric suctions (0.1, 33, and 1500 kPa, and 150 MPa) and Hungarian soil map
53 information. They found that the importance of soil properties in the prediction of the soil water
54 content varied according to soil type and matric suction. Recently Szabó et al. (2019) have
55 developed PTFs based on RF and geostatistics methods to map soil hydraulic properties, such as
56 water contents at saturation, field capacity and wilting point, for the Balaton catchment area in
57 Hungary. Araya and Ghezzehei (2019) compared the performances of four machine-learning
58 algorithms including the k-nearest neighbors (kNNs), support vector regression (SVR), RF, and
59 boosted regression tree (BRT) for prediction of saturated hydraulic conductivity. They found that
60 the BRT model outperformed the other algorithms closely followed by the RF model.
61 Gunarathna et al. (2019a) tested three machine-learning algorithms including ANN, kNN, and
62 RF to estimate volumetric water content at matric suctions of 10, 33 and 1500 kPa for soils in Sri
63 Lanka. They recommended that the PTFs to be developed using the RF algorithm. Ließ et al.
64 (2012) studied uncertainty in the spatial prediction of soil texture by comparison of the RF and
65 regression tree techniques for 56 soil profiles and found that the former method provided a better
66 result. Also, Wiesmeier et al. (2011) utilized the RF technique to develop digital mapping of the
67 soil organic matter content in 120 soil profiles. They found that the prediction accuracy of the RF
68 modeling was acceptable. A review of literatures therefore revealed that the RF data mining
69 technique has been applied to develop PTFs to predict specific points of the SWRC, such as field

70 capacity and permanent wilting point, or particular properties such as saturated hydraulic
71 conductivity, but it has not been used to develop parametric-based PTFs of the van Genuchten
72 model parameters, so far (Pages 5-6, lines 84-109).

73 Also, we have added new literatures, in which statistical and data mining techniques have been
74 used, to the introduction section of the manuscript, as follows:

75 Botula et al. (2013): Page 4, line 77.

76 Haghverdi et al. (2015): Page 4, line 77.

77 Nguyen et al. (2017): Page 4, line 78.

78 Gunarathna et al. (2019a): Page 4, line 74.

79 Gunarathna et al. (2019a): Page 4, line 77.

80 Gunarathna et al. (2019b): Page 4, line 72.

81 Khlosi et al. (2016): Page 4, line 78.

82 At page 6, lines 115-122 (section 2.2.), authors are presenting results in the Material and Methods
83 section. Therefore, this section should be moved to Results and Discussion section. By the way
84 the maximum clay content is 48% (Table 1), so the sentence should be rewritten accordingly.

85 Ans:

86 Following your suggestion, section 2.2 was moved and is now section 3.1 in the "Results and
87 discussion" section (Page 15, lines 283-293). Also, the sentence has been rewritten as follows:

88 "It can be seen that the average and maximum of clay content were 21.4 and 48%, respectively"

89 (Page 15, lines 285-286).

90 The same remark as above applies to page 7, lines 135-144 (section 2.4.): to move to Results and
91 Discussion section.

92 Ans:

93 Following your suggestion, section 2.4 was moved and is now section 3.2 in the "Results and
94 discussion" section (Pages 15-16, lines 294-331).

95 In addition, at line 138, authors are listing the soil properties that have high correlation with van
96 Genuchten parameters. They did not mention θ_{pWP} even it had high correlation coefficients!

97 Ans:

98 Thank you so much. It is a good point. This point has been mentioned in the manuscript.

99 Therefore, the sentence has been modified as follows:

100 "Clay and sand contents, θ_{FC} , θ_{pWP} , d_g and OM had the greatest significant correlations with the
101 parameters of the van Genuchten model (Fig. 4), which are consistent with other studies (Dexter
102 et al., 2008; Nemes et al., 2006). For example, the correlation coefficients between clay content
103 and θ_s ($r = 0.323$) is close to that between the OM and θ_s ($r = 0.268$). Also, the results showed that
104 there were significant correlations between θ_{pWP} and input variables of clay content (+), sand
105 content (-), BD (-), OM (+) and K_s (-), and also between θ_{pWP} and θ_s (+) and n (-) parameters of
106 the van Genuchten model (Fig. 4) (Fig. 4). Botula et al. (2012) also found the same observation
107 for the correlation of θ_{pWP} with sand and clay contents and BD of tropical Lower Congo soils (Page
108 16, lines 299-307).

109 Also, at lines 143 and 144, the authors stated that there was no correlation between van Genuchten
110 parameters and K_s whereas they used this soil property in PTF14 and PTF15. Could they explain
111 why they considered K_s even if it not correlated to van Genuchten parameters?

112 Ans:

113 There are many cases, where two variables might not show a strong simple correlation, but may
114 show a strong association in the regression, along with other predictors. In other words, the simple
115 correlation coefficient is a way to show the relationship between independent and dependent

116 variables, but it cannot show a model for the relationship between these two variables, when other
117 independent variables have been used in a multiple regression (Simmons et al., 2011). The result
118 of multiple regression analysis with backward selection method showed that the K_s variable
119 remained in the PTF14 and PTF15 for all the van Genuchten model parameters. Some of the
120 regression equations with backward selection method are shown in the following as examples:

121 $\theta_r = -0.69 + 0.22 \times \text{Clay} + 0.278 \times \text{Sand} + 0.20 \times K_s, R = 0.31^{**}$

122 $\alpha = -3.72 + 0.23 \times \text{Clay} + 0.17 \times \text{BD} + 0.282 \times K_s, R = 0.33^{**}$ and

123 $n = -1.76 + 0.24 \times \text{Sand} + 0.164 \times K_s, R = 0.30^{**}.$

124 On the other hand, the non-linear correlations between variables are very important in this study.
125 Both the multiple NLR approach and RF data mining technique are non-linear prediction methods.
126 Fig. 4 only shows simple linear correlation between variables, but there may be non-linear
127 correlations between variables, which may affect the estimation of the dependent variables. For
128 example, the results of non-linear correlations showed that K_s had strong correlations with θ_s and
129 α of the van Genuchten model parameters by logarithmic ($\theta_s = 0.652 - 0.027 \times \ln K_s, R = 0.62^{**}$) and
130 power ($\alpha = 0.007 \times K_s^{0.283}, R = 0.57^{**}$) equations, respectively, which were greater than their simple
131 correlations (Pages 16-17, lines 310-328). In support of this claim, the results showed that by
132 adding OM and/or K_s as predictors in the PTFs 13, 14 and 15, the accuracy (Fig. 5B) and reliability
133 (Fig. 6B) of the prediction of the SWRC improved by 16, 13, 17 and 7.1, 6.3, 6.9%, respectively,
134 compared to the PTF3 in terms of the *IRMSE* criterion in the RF method (Pages 25-26, lines 517-
135 520).

136 At page 8, line 152, the authors assessed multicollinearity using the variance inflation factor (VIF)
137 in the Material and Methods section but they reported nothing about this in the Results and

138 Discussion section; although they mentioned that silt content was eliminated to avoid
 139 multicollinearity (line 164).

140 Ans:

141 You are completely right. The values of variance inflation factor (*VIF*) for all PTFs have been
 142 added to the manuscript. Therefore, the text has been modified as follows:

143 "Before developing PTFs, all variables were evaluated by Kolmogorov-Smirnov normality and
 144 multicollinearity tests by the SPSS 24 software (IBM, 2016). The degree of multicollinearity in
 145 the PTFs was tested by the variance inflation factor ($VIF=1/1-R^2_j$, where R^2_j is the R^2 value
 146 obtained by regressing the j^{th} predictor on the remaining predictors) (Hocking, 2013). Also, to
 147 avoid multicollinearity between textural contents, the silt fraction was not used as a predictor"
 148 (Page 9, lines 157-161). Results of the multicollinearity analysis (*VIF*) are shown in Table 3. The
 149 *VIF* values showed low levels of multicollinearity among the independent variables ($VIF<10$)
 150 (Khodaverdiloo et al., 2011) (Page 17, lines 334-336).

151 **Table 3-** The variance inflation factor (*VIF*) values for normalized form of input variables.

PTFs	Clay* (%)	Sand (%)	BD ^s (g cm ⁻³)	θ_{FC} (cm ³ cm ⁻³)	θ_{PWP} (cm ³ cm ⁻³)	d_g (mm)	δ_g (-)	TP (cm ³ cm ⁻³)	OM (%)	K_s (cm day ⁻¹)
PTF2	1.42	1.42								
PTF3	1.43	1.52	1.10							
PTF4	1.45	1.56	1.25	1.31						
PTF5	1.79	1.58	1.27	2.48	2.56					
PTF6						1.00	1.00			
PTF7			1.11			1.11	1.01			
PTF8			1.25	1.33		1.01	1.22			
PTF9			1.28	2.50	2.73	1.34	1.22			

PTFs	Clay* (%)	Sand (%)	BD [§] (g cm ⁻³)	θ_{FC} (cm ³ cm ⁻³)	θ_{PWP} (cm ³ cm ⁻³)	d_g (mm)	δ_g (-)	TP (cm ³ cm ⁻³)	OM (%)	K _s (cm day ⁻¹)
PTF10	1.55	1.43						1.11		
PTF11	1.58	1.46		1.32				1.26		
PTF12	1.60	1.79		2.49	2.56			1.28		
PTF13	1.48	1.65	1.25						1.14	
PTF14	1.55	1.64	1.14							1.06
PTF15	1.55	1.65	1.25						1.15	1.06

152 * Normalized form of input variables is available in Table 2.

153 §. A list of abbreviations is available in the notation box.

154 Page 10, lines 198-203, the authors used 20-fold cross validation: the question why the authors
 155 used this specific k value and not, for example, 10 which is the most commonly used one in cross
 156 validation?

157 Ans:

158 In the present study, the k-fold cross validation approach (Efron and Tibshirani, 1994) was
 159 utilized to obtain training and testing datasets for each PTF. The number of folds (i. e., k) was
 160 obtained by trial and error. To do so, some PTFs, selected randomly, were developed with 10, 15
 161 and 20-fold cross-validation. Then, the k value which resulted in the best performance of the
 162 PTFs, was selected to develop all PTFs in this study. The results showed that 20-fold cross
 163 validation performed better than the other folds in most of the PTFs (Table 1). Therefore, 20-fold
 164 cross validation was selected to develop PTFs in this study (page 11, lines 201-207).

165 **Table 1-** The results of 10, 15 and 20-fold cross-validation (k) for van Genuchten model
 166 parameters of the soil water retention curve derived from nonlinear regression (NLR) and

167 random forest (RF) techniques based on root mean square error (*RMSE*) for pedotransfer
 168 functions PTF 3, 5 and 11 in the train and test datasets.

			θ_r			θ_s			α			n		
			<i>RMSE</i>			<i>RMSE</i>			<i>RMSE</i>			<i>RMSE</i>		
			Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	Train	Test	Mean
PTF3	k=10	NLR	0.058	0.060	0.059	0.063	0.065	0.064	1.017	1.037	1.027	0.426	0.436	0.431
		RF	0.052	0.061	0.056	0.058	0.073	0.066	0.893	1.084	0.989	0.374	0.442	0.408
	k=15	NLR	0.058	0.060	0.059	0.064	0.064	0.064	1.017	1.030	1.024	0.426	0.434	0.430
		RF	0.052	0.061	0.057	0.058	0.070	0.064	0.894	1.033	0.964	0.374	0.441	0.408
	k=20	NLR	0.058	0.060	0.059	0.064	0.064	0.064	0.064	0.064	0.064	0.426	0.437	0.432
		RF	0.051	0.060	0.056	0.057	0.071	0.064	0.057	0.071	0.064	0.368	0.442	0.405
PTF5	k=10	NLR	0.051	0.053	0.052	0.053	0.054	0.054	0.764	0.796	0.780	0.380	0.397	0.389
		RF	0.043	0.056	0.050	0.046	0.056	0.051	0.675	0.869	0.772	0.327	0.411	0.369
	k=15	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.764	0.790	0.777	0.381	0.399	0.390
		RF	0.044	0.054	0.049	0.046	0.055	0.050	0.679	0.848	0.763	0.329	0.421	0.375
	k=20	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.765	0.789	0.777	0.381	0.399	0.390
		RF	0.042	0.054	0.048	0.044	0.054	0.049	0.654	0.842	0.748	0.316	0.412	0.364
PTF11	k=10	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.018	1.052	1.035	0.431	0.448	0.440
		RF	0.050	0.061	0.056	0.047	0.057	0.052	0.770	0.978	0.874	0.370	0.443	0.406
	k=15	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.019	1.037	1.028	0.432	0.447	0.439
		RF	0.050	0.060	0.055	0.047	0.057	0.052	0.770	1.009	0.889	0.369	0.450	0.410
	k=20	NLR	0.058	0.060	0.059	0.065	0.065	0.065	1.020	1.024	1.022	0.432	0.439	0.435
		RF	0.049	0.061	0.055	0.046	0.056	0.051	0.745	0.964	0.855	0.361	0.443	0.402

169
 170 Also, the authors used data from 6 different provinces and 2 soil depths. I wonder if they took
 171 into consideration these two distinguishing factors when they split their data during k-fold cross
 172 validation into training and testing subsets.

173 **Ans:**

174 All soil samples, which have been collected from 6 different provinces and 2 soil depths, have
 175 been assumed as one database and training and testing data have been selected randomly from the
 176 database (which have been included all soil samples). In other words, we have not taken into
 177 consideration these two distinguishing factors (province or depth of sampling) when we split all
 178 data during k-fold cross validation into training and testing subsets.

179 Page 10, line 208 and equation (2): the authors noted the number of input variables by n ; there
180 may be confusion with the fourth parameter of van Genuchten model (page 7, equation (1) and
181 line 131)! Here n may be replaced by p (the number of input variables like in the AIC definition
182 at page 13, equation (6). By the way the authors should use the same letter: p and not P (line 256)!

183 [Ans:](#)

184 [Thank you so much. The required correction has been done \(Page 11, line 216\).](#)

185 Page 13, lines 258-260: the average values can be compared using the analysis of variance
186 (ANOVA) method and, once they are significantly different, we can use some posthoc tests like
187 the Duncan test. However, it is not clear what was compared: all the 15 PTFs for both RF and
188 NLR, and even from Rosetta for the testing datasets (Figures 6 and 7, graphs B) or the 2 or 3
189 methods (NLR, RF, and Rosetta) separately for each of the 15 PTFs (page 14, lines 270-273). If it
190 is the former case, Duncan test is useless since it compares 30 mean values (and even 35 if we
191 consider Rosetta in addition to NLR and RF) and consequently some PTFs are belonging to 2 or 3
192 different groups (like PTF4 RF, PTF5 NLR, etc. with abc letters) for training data sets (Figure 6)
193 and even more for the testing dataset (4 letters like h-k or i-l on Figure 7). Moreover, this statistical
194 comparison was done only for IRMSE but not for the 3 other cross validation criteria (IME, R^2 ,
195 and AIC). Is there any explanation?

196 [Ans:](#)

197 [Due to the fact that the performance of both methods was evaluated for all samples, therefore the](#)
198 [mean comparison test can be used to compare the predicting accuracy and reliability of the RF and](#)
199 [NLR methods. In other words, to determine whether the differences in the accuracy and reliability](#)
200 [of the RF and NLR methods are random or real, the mean comparison test could be performed.](#)
201 [One of the aims of the present study was to investigate the performance of each method in different](#)

202 PTFs. In other words, the performance of each method in each PTF was important to the users.
203 "To evaluate the performance of each method in different PTFs, the effect of method as the first
204 factor at two levels in the training step (*i.e.*, NLR and RF methods) and at three levels in the testing
205 step (*i.e.*, NLR, RF and Rosetta methods), and the different PTFs as the second factor at 15 levels
206 (PTF1 to PTF15), were investigated using a two-way analysis of variance (ANOVA) with a
207 randomized complete block design, based on the *IRMSE* of prediction of the SWRC" (Pages 14-
208 15, lines 270-275). Table 4 shows the results of the ANOVA of the *IRMSE* of prediction of the
209 SWRC by different methods and PTFs. The effect of methods and PTFs, and their interaction, on
210 the *IRMSE* was significant at $P < 0.01$, 0.01 and 0.05, respectively, in the training step, and at
211 $P < 0.01$, 0.01 and 0.01, respectively, in the testing step. Therefore, we focus on the results and
212 discussion of the comparison of the method \times PTF interaction effects (Page 18, lines 340-346).
213 The *IRMSE* criterion calculates the total error, including bias and random errors, and is a more
214 appropriate criterion for evaluating the accuracy and reliability of the RF and NLR methods
215 compared to other criteria (Chai and Draxler, 2014). Therefore, to compare the predicting accuracy
216 and reliability of the RF and NLR methods, the average values of the *IRMSE* was compared with
217 Duncan's test by MathWorks (2018) software (Page 15, lines 275-280).
218 **Table 4-** Analysis of variance of the integral root mean square error (*IRMSE*) of the prediction of
219 the soil water retention curve by different methods (nonlinear regression and random forest) and
220 pedotransfer functions (PTFs 1-15) for both the train and test datasets.

	Source	Degree freedom	Mean square	F-value	P-value
Train	Repeat (Block)	222	0.007	19.09	<0.0001
	PTFs	14	0.062	180.68	<0.0001
	Methods	1	0.038	109.69	<0.0001
	PTFs \times Methods	14	0.001	1.78	0.0356
	Error	6288	0.0003		
Test	Repeat (Block)	222	0.010	16.04	<0.0001
	PTFs	14	0.073	117.22	<0.0001
	Methods	2	0.656	1056.43	<0.0001

PTFs × Methods	18	0.002	3.68	<0.0001
Error	7398	0.0006		

221

222 At page 19, lines 385-387: the authors are discussing the correlation between thetar and referring
 223 to Figure 2 whereas correlation coefficients between thetar and soil proprieties were not included
 224 in this figure!

225 **Ans:**

226 The correlation test was not performed for the θ_r variable, because its value was zero in 138 out of
 227 223 soil samples, as has been reported in other studies (Campbell and Horton Jr, 2002; Rawls et
 228 al., 1991; Tomasella et al., 2000) for θ_r variable (Pages 15-16, lines 296-299). Therefore, the
 229 sentence has been rewritten as follows: "Therefore, input variables of the textural contents or
 230 statistics can influence the residual saturation region of the SWRC. However, soil water content at
 231 the dry end (high matric suctions) of the SWRC is primarily determined by textural contents
 232 (Hillel, 1998) " (Pages 23-24, lines 470-473).

233 In addition, the whole subsection 3.1.2.2. is about the importance of the introduction of Ks into
 234 PTF 14 and 15 whereas there was no correlation between van Genuchten parameters and Ks.
 235 How the authors can explain the added value of Ks to the last 2 PTFs even in the absence of
 236 significant correlation?

237 **Ans:**

238 It has been answered in pages 5-6, lines 113-135 of this file.

239 Furthermore, at page 21, lines 442 and 443, the authors said that Ks is correlated to soil texture
 240 and TP variables whereas it is correlated only to clay content (Figure 2) but not to sand nor to TP.

241 **Ans:**

242 Thank you so much. The sentence has been rewritten as follows:

243 "The correlation results showed (Fig. 4) that K_s can be strongly influenced by clay content and
244 textural statistics (d_g and δ_g)" (Page 26, lines 524-525).

245

246 **-Reviewer 2**

247 -I thank the authors for their through addressing my queries and completing the recommended
248 revisions. The authors should address following points.

249 **Ans:**

250 Thank you so much. Your comments helped us a lot to improve the manuscript.

251 1. Revise L45-46 as follows: "These findings could provide the scientific basis for further
252 research on the RF method."

253 **Ans:**

254 It has been done (page 2, lines 45-46).

255 2. I could not find the following revision in the revised manuscript, please recheck for its
256 existence.

257 L104-105: What do you mean by "topsoil" and "subsoil"? Do you mean A and B horizons or
258 tillage depth? Be specific. Also, what do you mean with layer in "depending on thickness of
259 layers"?

260 **Ans:**

261 "topsoil" and "subsoil" refer to A and B horizons, respectively. It was corrected in the
262 manuscript. Since the sampling was done from different locations of the various provinces, the
263 topsoil and subsoil layers of soil at different locations had different depths and thicknesses, and
264 samples were taken from the center of each layer. Therefore, the samples were taken from different
265 depths, depending on thickness of the A and B layers.

266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

Ans:

Thank you so much. It has been arranged as follows:

Since the sampling was done from different locations of the various provinces, the topsoil and subsoil layers of soil at different locations had different depths and thicknesses. We collected samples from the center of the topsoil and subsoil layers, which represented the pedological A and B horizons, respectively. The sampling depths varied from 10 to 35 cm for topsoil (208 samples) and from 20 to 45 cm for subsoil (15 samples), reflecting the variation in the soil profiles (pages 6-7, lines 123-130).

3. I do recommend the authors go over the manuscript for mistakes of grammar, typos, sentence structure, and so on before sending their final copy to the editor.

Ans:

We thank the reviewer for this point. The co-author for whom English is their first language has been through the manuscript thoroughly and corrected all errors in spelling and grammar.

Eventually;
As it was described point by point, the manuscript was revised significantly.

Acknowledgements

The authors are deeply grateful to anonymous reviewers and the editor for their helpful comments on the manuscript.

291 **References**

- 292 Araya, S.N., Ghezzehei, T.A., 2019. Using Machine Learning for Prediction of Saturated
293 Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. *Water Resour.*
294 *Res.* 55(7), 5715-5737.
- 295 Botula, Y.-D., Cornelis, W., Baert, G., Van Ranst, E., 2012. Evaluation of pedotransfer functions
296 for predicting water retention of soils in Lower Congo (DR Congo). *Agric. Water Manag.*
297 111, 1-10.
- 298 Botula, Y.-D., Cornelis, W.M., Baert, G., Mafuka, P., Van Ranst, E., 2013. Particle size
299 distribution models for soils of the humid tropics. *Journal of Soils and Sediments* 13(4),
300 686-698.
- 301 Campbell, G.S., Horton Jr, R., 2002. *Methods of Soil Analysis: Part 4, Physical Methods.* Soil
302 *Sci. Soc. Am.*
- 303 Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–
304 Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7(3), 1247-
305 1250.
- 306 Dexter, A., Czyż, E., Richard, G., Reszkowska, A., 2008. A user-friendly water retention
307 function that takes account of the textural and structural pore spaces in soil. *Geoderma*
308 143(3), 243-253.
- 309 Efron, B., Tibshirani, R.J., 1994. *An introduction to the bootstrap.* CRC press.
- 310 Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., 2019a. Machine Learning
311 Approaches to Develop Pedotransfer Functions for Tropical Sri Lankan Soils. *Water*
312 11(9), 1940.

313 Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., Amarasekara, M., 2019b.
314 Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. *Soil*
315 *Till. Res.* 190, 109-119.

316 Haghverdi, A., Leib, B.G., Cornelis, W.M., 2015. A simple nearest-neighbor technique to predict
317 the soil water retention curve. *Transactions of the ASABE* 58(3), 697-705.

318 Hillel, D., 1998. *Environmental soil physics: Fundamentals, applications, and environmental*
319 *considerations.* Academic press.

320 Hocking, R.R., 2013. *Methods and applications of linear models: regression and the analysis of*
321 *variance.* John Wiley & Sons.

322 IBM, C., 2016. *IBM SPSS Statistics for Windows, Version 24.0.* Armonk, NY: IBM Corp.

323 Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced
324 pedotransfer functions with support vector machines to predict water retention of
325 calcareous soil. *Eur. J. Soil Sci.* 67(3), 276-284.

326 Khodaverdiloo, H., Homae, M., van Genuchten, M.T., Dashtaki, S.G., 2011. Deriving and
327 validating pedotransfer functions for some calcareous soils. *J. Hydrol.* 399(1), 93-99.

328 Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture:
329 comparison of regression tree and Random Forest models. *Geoderma* 170, 70-79.

330 MathWorks, 2018. *MATLAB: the language of technical computing.* Inc., Natick, Massachusetts,
331 United States.

332 Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor
333 approach to estimate soil hydraulic properties. *Soil Sci. Soc. Am. J.* 70(2), 327-336.

334 Nguyen, P.M., Haghverdi, A., De Pue, J., Botula, Y.-D., Le, K.V., Waegeman, W., Cornelis,
335 W.M., 2017. Comparison of statistical regression and data-mining techniques in
336 estimating soil water retention of tropical delta soils. *Biosyst. Eng.* 153, 12-27.

337 Rawls, W., Gish, T., Brakensiek, D., 1991. Estimating soil water retention from soil physical
338 properties and characteristics, *Advances in soil science*. Springer, pp. 213-234.

339 Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed
340 flexibility in data collection and analysis allows presenting anything as significant.
341 *Psychol. Sci.* 22(11), 1359-1366.

342 Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019.
343 Mapping soil hydraulic properties using random forest based pedotransfer functions and
344 geostatistics. *Hydrol. Earth Syst. Sci.* 23(6), 2615-2635.

345 Tomasella, J., Hodnett, M.G., Rossato, L., 2000. Pedotransfer functions for the estimation of soil
346 water retention in Brazilian soils. *Soil Sci. Soc. Am. J.* 64, 327-338.

347 Tóth, B., Makó, A., Toth, G., 2014. Role of soil properties in water retention characteristics of
348 main Hungarian soil types. *J. Cent. Eur. Agric.* 15(2), 137-153.

349 Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic
350 matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant Soil*
351 340(1), 7-24.

352

1 **Estimating the soil water retention curve: comparison of multiple nonlinear regression**
2 **approach and random forest data mining technique**

3 **M. Rastgou¹, H. Bayat^{2*}, and M. Mansoorizadeh³, Andrew S. Gregory⁴**

4
5 ¹ **Mostafa Rastgou:** Ph. D. Student of Soil Science, Department of Soil Science, Faculty of
6 Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: mostafa.rastgo@gmail.com,

7 ² **Hossein Bayat:** Associate Professor (Ph. D.), Department of Soil Science, Faculty of Agriculture,
8 Bu Ali Sina University, Hamedan, Iran. Postal Address: Department of Soil Science, Faculty of
9 Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: h.bayat@basu.ac.ir Other e-mail:
10 hbayat2001@gmail.com. Office phone: +98-81-34424189, Mobile phone: +98-918-8188378.
11 Fax: +98-81-34424189.

12 ³ **Muharram Mansoorizadeh:** Assistant professor (Ph. D.), Department of Computer Science,
13 Faculty of Engineering, Bu Ali Sina University, Hamedan, Iran. E-mail: mansoorm@basu.ac.ir

14 ⁴ **Andrew S. Gregory:** Sustainable Agriculture Sciences Department, Rothamsted Research,
15 Harpenden, Hertfordshire, AL5 2JQ, UK. E-mail: andy.gregory@rothamsted.ac.uk

16
17 *Corresponding author (h.bayat@basu.ac.ir, Other e-mail: hbayat2001@gmail.com).

24 **Estimating the soil water retention curve: comparison of multiple nonlinear regression**
25 **approach and random forest data mining technique**

26 **Abstract**

27 This study evaluates the performance of the random forest (RF) method on the prediction of the
28 soil water retention curve (SWRC) and compares its performance with those of non-linear
29 regression (NLR) and Rosetta-based pedotransfer functions (PTFs), which has not been reported
30 so far. Fifteen RF and NLR-based PTFs were constructed using readily-available soil properties
31 for 223 soil samples from Iran. The general performance of RF and NLR-based PTFs was
32 quantified by the integral root mean square error (*IRMSE*), Akaike's information criterion (*AIC*)
33 and coefficient of determination (R^2). The results showed that the accuracy of the RF-based PTFs
34 was significantly ($P < 0.05$) better than the NLR-based PTFs, and ~~also, that~~ the reliability of the
35 NLR-based PTFs was significantly ($P < 0.01$) better than the RF-based PTFs and all of the
36 Rosetta-based PTFs. The average values of the *IRMSE*, *AIC* and R^2 of the RF method were 0.041
37 $\text{cm}^3 \text{cm}^{-3}$, -16997.7, ~~and~~ 0.987, and 0.053 $\text{cm}^3 \text{cm}^{-3}$, -15547.5, ~~and~~ 0.981 for the training and
38 testing steps of all PTFs, respectively, whereas ~~these~~ values for the NLR method were 0.046 cm^3
39 cm^{-3} , -16616.4, ~~and~~ 0.984, and 0.048 $\text{cm}^3 \text{cm}^{-3}$, -16355.6, ~~and~~ 0.983 for the training and testing
40 steps, respectively. The PTF5 of the RF and NLR methods, with ~~the~~ inputs of sand and clay
41 contents, bulk density, and the water content at field capacity and permanent wilting point, had
42 the greatest R^2 values (0.987 and 0.989, ~~respectively~~), and the lowest *IRMSE* values (0.039 and
43 0.032 $\text{cm}^3 \text{cm}^{-3}$, ~~respectively~~), ~~respectively~~, compared to other PTFs for the testing step. Overall,
44 the RF method had less reliability for the prediction of the SWRC compared to the NLR method
45 due to overprediction, uncertainty of determination of forest scale and instability in the testing

46 step. It seems that these findings could provide the scientific basis for further research on the
47 RF method.

48 *Keywords:* pedotransfer functions; soil water retention curve; soil texture; soil structure; van
49 Genuchten.

50

Notation	
Sand content (%)	S
Clay content (%)	C
Geometric mean diameter (mm)	d_g
Geometric standard deviation (-)	δ_g
Bulk density (g cm^{-3})	BD
Total porosity ($\text{cm}^3 \text{cm}^{-3}$)	TP
Water content at field capacity, 33 kPa ($\text{cm}^3 \text{cm}^{-3}$)	θ_{FC}
Water content at 1500 kPa ($\text{cm}^3 \text{cm}^{-3}$)	θ_{PWP}
Organic matter content (%)	OM
Saturated hydraulic conductivity (cm day^{-1})	K_s
Saturated water content ($\text{cm}^3 \text{cm}^{-3}$)	θ_s
Residual water content ($\text{cm}^3 \text{cm}^{-3}$)	θ_r
Random forest	RF
Nonlinear regression	NLR
Soil water retention curve	SWRC

51

52 **1 Introduction**

53 Soil hydraulic properties are principle factors that control the movement of water and solutes in
54 the soil. Determination of the soil hydraulic properties is required for many distinct applications
55 linked with irrigation, land use planning, drainage and drought risk assessment (Dobarco et al.,
56 2019). The soil water retention curve (SWRC) is one of the most important soil hydraulic

57 properties. It defines the relationship between soil matric potential and soil water content (Hillel,
58 1998). The SWRC is a crucial parameter in soil and water management for sustainable and
59 improved agricultural production (Shwetha and Varija, 2015). The SWRC depends principally
60 on texture, structure and bulk density (BD) of soils (Wassar et al., 2016). Many methods have
61 been introduced for the direct measurement of the SWRC in the laboratory (e.g., the hanging
62 water column and pressure plate methods) (Klute, 1986) and in the field (e.g., tensiometric)
63 (Bruce and Luxmoore, 1986). Measurements of the SWRC at several matric potentials can be
64 expensive, difficult and time-consuming, hence it is common to predict it by modelling (Dobarco
65 et al., 2019). Modelling of soil water is an essential tool in evaluating the effects of different
66 managements on crop yield and environmental quality (Verhagen, 1997).

67 Pedotransfer functions (PTFs) translate easy-to-measure data that we have (e.g., texture class,
68 particle size distribution (PSD) and BD) into difficult-to-measure data that we need (soil
69 hydraulic data) (Bouma, 1989). Estimates of the SWRC by the-PTFs are valuable in many
70 studies, such as hydrology, soil mapping and hydrogeology (Børgesen and Schaap, 2005). The
71 point- and parametric-based PTFs are generally developed to predict water content at certain
72 specific matric potential values and the entire SWRC, respectively, by multiple linear (MLR) and
73 nonlinear regression (NLR) methods (Gunarathna et al., 2019b; Merdun et al., 2006; Minasny et
74 al., 1999; Rajkai et al., 2004; Tomasella et al., 2000). Data mining techniques including artificial
75 neural networks (ANNs) (Bayat et al., 2013a; Bayat et al., 2013b; Gunarathna et al., 2019a;
76 Koekkoek and Booltink, 1999; Pachepsky et al., 1996), group method of data handling (GMDH)
77 (Bayat et al., 2011; Neyshaburi et al., 2015; Pachepsky and Rawls, 1999), nonparametric nearest
78 neighbor technique (Botula et al., 2013; Gunarathna et al., 2019a; Haghverdi et al., 2015; Nemes
79 et al., 2006; Nguyen et al., 2017) and support vector machine (SVM) (Khlosi et al., 2016;

80 Lamorski et al., 2008; Lamorski et al., 2014; Twarakavi et al., 2009), have been applied
81 successfully ~~applied~~ for PTF development.

82 Random forest (RF), or random decision forests, has become a popular approach as an ensemble
83 learning method for prediction and classification (Verikas et al., 2011). The RF method has been
84 developed by Breiman (2001) as an expansion of the classification and regression trees (CART)
85 technique to provide better performance of prediction results (Wiesmeier et al., 2011). So far,
86 few studies have been carried out on the application of the RF method in soil science. ~~For~~
87 ~~example,~~ Tóth et al. (2014) applied the RF method to analyze the relationship between soil water
88 content at four matric suctions of (0.1, 33, and 1500 kPa, and 150000 kMPa) and Hungarian soil
89 map information. They found that the importance of soil properties in the prediction of the soil
90 water content varied, according to soil type and matric suctions. Recently Szabó et al. (2019)
91 have developed PTFs based on RF and geostatistics methods to map soil hydraulic properties,
92 such as water contents at saturation, field capacity and wilting point, for the Balaton catchment
93 area in Hungary. Araya and Ghezzehei (2019) compared the performances of four machine-
94 learning algorithms including -the k-nearest neighbors (kNNs), support vector regression (SVR),
95 RF, and boosted regression tree (BRT) for prediction of the saturated hydraulic conductivity.
96 They found that the BRT models outperformed the other algorithms closely followed by the RF
97 models. Gunarathna et al. (2019a) tested three machine-learning algorithms including artificial
98 neural networks (ANN), kNN, and RF to estimate volumetric water content at the matric suctions
99 of 10, 33 and 1500 kPa for soils in Sri Lankan soils. They recommended that the PTFs to be
100 developed using the RF algorithm. Ließ et al. (2012) studied uncertainty in the spatial prediction
101 of soil texture by comparison of the RF and regression tree techniques for 56 soil profiles. Those
102 authors indicated and found that the RF former method provided a better results better than the

103 ~~regression tree~~. Also, Wiesmeier et al. (2011) utilized the RF technique to develop digital
104 mapping of the soil organic matter content in 120 soil profiles. They ~~pointed-out~~found that the
105 prediction accuracy of the RF modeling was acceptable. A review of literatures therefore
106 revealed that the RF data mining technique has been ~~only~~ applied to develop PTFs to predict
107 specific points-based PTFs of the SWRC, such as ~~including~~ field capacity and permanent wilting
108 point, or particular properties such as saturated hydraulic conductivity, but it has not been used
109 for to developing parametric-based PTFs of the van Genuchten model parameters, so far~~The RF~~
110 ~~data mining technique has not been applied to predict the SWRC, so far~~. Therefore, the objective
111 of the present study was to develop simple parametric-PTFs to predict the SWRC with greater
112 accuracy and reliability using a novel approach with the RF data mining technique. ~~We and~~
113 compare its performance with those of the multiple ~~non-linear regression (NLR)~~ approach and
114 with Rosetta software (Schaap et al., 2001) on the prediction of the SWRC through finding the
115 best input variables and PTFs for the SWRC.

116

117 **2 Materials and methods**

118 *2.1 Sample collection and determination*

119 In the present study 223 undisturbed and disturbed soil samples were taken from six provinces of
120 Iran including west Azarbaijan ($35^{\circ} 8' - 39^{\circ} 46' N, 44^{\circ} 3' - 47^{\circ} 23' E$; 60 data), Hamedan
121 ($33^{\circ} 59' - 35^{\circ} 48' N, 47^{\circ} 34' - 49^{\circ} 36' E$; 55 data), Kermanshah ($33^{\circ} 41' - 35^{\circ} 17' N,$
122 $45^{\circ} 24' - 48^{\circ} 6' E$; 26 data), Kurdistan ($34^{\circ} 45' - 36^{\circ} 31' N, 45^{\circ} 31' - 48^{\circ} 13' E$; 22
123 data), Mazandaran ($35^{\circ} 46' - 36^{\circ} 58' N, 50^{\circ} 21' - 58^{\circ} 08' E$; 30 data) and Fars ($27^{\circ} 2' -$
124 $31^{\circ} 42' N, 50^{\circ} 42' - 55^{\circ} 38' E$; 30 data). Steel cylinders, measuring 5.1 cm in diameter and
125 3.5 cm in height, were used to collect the undisturbed samples. Since the sampling was done

126 from different locations of the various provinces, the topsoil and subsoil layers of soil at different
127 locations had different depths and thicknesses. ~~We collected, and samples were taken~~ from the
128 center of ~~the topsoil and subsoil each~~ layers, which represented (~~"topsoil" and "subsoil" refer~~
129 ~~to the pedological A and B horizons, respectively~~). ~~Therefore, the samples were taken from~~
130 ~~different depths, depending on the thickness of the A and B layers~~. The sampling depths varied
131 from 10 to 35 cm for topsoil (~~A horizon~~, 208 samples) and from 20 to 45 cm for subsoil (~~B~~
132 ~~horizon~~, 15 samples), ~~reflecting the variation in the soil profiles~~.
133 Soil PSD was analyzed by ~~the~~ hydrometer method (Gee and Or, 2002), and the geometric mean
134 and standard deviation of particles diameter (d_g and δ_g , respectively) were calculated by
135 equations from Shirazi and Boersma (1984). Organic matter (OM) content was determined by
136 the Walkley and Black (1934) method and BD by the core method (Blake and Hartge, 1986).
137 Total porosity (TP) was calculated from BD and particle density, and the saturated hydraulic
138 conductivity (K_s) was measured with a constant head permeameter (Klute and Dirksen, 1986).
139 The SWRC was ~~constructed~~ ~~constructed~~ by measuring the volumetric water content at ~~the~~
140 matric suctions of 0 (saturation status of soil samples), 1, 2 and ~~5 kPa~~ ~~5 kPa~~ with a sandbox
141 apparatus, and at 10, 25, 50, 100, 200, 500, 1000 and 1500 kPa with a pressure plate apparatus.
142 Undisturbed samples were used for measurement of the matric suctions from 0 to 100 kPa and
143 disturbed samples were used for matric suctions from 200 to 1500 kPa. Two key points in the
144 SWRC are the water contents at field capacity (30 kPa suction; θ_{FC}) and permanent wilting point
145 (1500 kPa suction; θ_{PWP}).

146

147 2.2 Soil-water retention equation

148 The van Genuchten–Mualem (Eq. (1)) model (Mualem, 1976; van Genuchten, 1980) was utilized
149 to describe the SWRC data.

$$\theta = \theta_r + (\theta_s - \theta_r) \times \frac{1}{\left[1 + (\alpha h)^n\right]^{\left(\frac{1}{n}\right)}} \quad (1)$$

150 where θ_r and θ_s are residual and saturated water contents ($\text{cm}^3 \text{ cm}^{-3}$), respectively, and h is the
151 soil water suction (kPa). The parameter α is related to the inverse of the air entry pressure (>0 ,
152 kPa^{-1}) and n (>1 , dimensionless parameter) is related to the pore size distribution of the soil (van
153 Genuchten, 1980). In the present study, van Genuchten model parameters θ_r , θ_s , α and n were
154 obtained using the MATLAB software (MathWorks, 2018).

155

156 2.3 Data pre-processing

157 Data pre-processing and regression assumptions, including detection of outliers, normality test of
158 the residuals, multicollinearity and independence of the residuals, were applied for all variables
159 (Berry, 1993). The outliers in the data were identified by the inter-quartile range (IQR) method
160 (Seo, 2006) and were replaced by the lower and upper threshold values (MathWorks, 2018).
161 Before developing PTFs, all variables were evaluated by Kolmogorov-Smirnov normality and
162 multicollinearity tests by the SPSS 24 software (IBM, 2016). The degree of multicollinearity in
163 the PTFs was tested by the variance inflation factor ($VIF=1/1-R_j^2$, where R_j^2 is the R^2 value
164 obtained by regressing the j^{th} predictor on the remaining predictors) (Hocking, 2013) (Table 1).
165 The VIF values in Table 1 showed low levels of multicollinearity among the independent
166 variables ($VIF<10$) (Khodaverdilo et al., 2011). Also, to avoid multicollinearity between
167 textural contents, the silt fraction was eliminated not used as a predictor. The variables clay
168 content, sand content, d_g , δ_g , OM, K_s , α and n had non-normal distributions, therefore,
169 transformations were applied to normalize them.

170

171 2.4 Developing PTFs

172 The PTF inputs were arranged in four steps (Fig. 21). The first step (PTFs 1-5) was based on
173 basic soil properties (i.e., sand content (%), clay content (%), BD (g cm^{-3}), θ_{FC} ($\text{cm}^3 \text{cm}^{-3}$) and
174 θ_{PWP} ($\text{cm}^3 \text{cm}^{-3}$)) according to Rosetta-based PTFs (Schaap et al., 2001) for comparison of
175 SWRC estimates by other methods. ~~To avoid multicollinearity between textural contents, the silt~~
176 ~~fraction was eliminated.~~ The parameters of the van Genuchten model were predicted in all steps.
177 In the second step (PTFs 6-9), d_g (mm) and δ_g were used as new inputs instead of sand and clay
178 contents in the previous step to evaluate the efficiency of using statistical descriptors of PSD to

179 predict the parameters of the van Genuchten model. To build the third step (PTFs 10-12), TP
180 ($\text{cm}^3 \text{ cm}^{-3}$) replaced BD from PTFs 3-5 to evaluate the effect of using TP instead of BD on the
181 prediction of the parameters of the van Genuchten model. In other words, the purpose of the
182 second and third steps was to evaluate whether the use of another form of descriptors of ~~the~~ soil
183 structure (TP instead of the BD) and soil texture (d_g and δ_g instead of the sand and clay contents)
184 would improve the accuracy of the estimates or not. In the last step, PTFs 13-15 were developed
185 by including OM (%) and K_s (cm day^{-1}) as new variables to evaluate the efficiency of these
186 instead of the water content at specific matric suctions on the prediction of the van Genuchten
187 model parameters. The input variables of the 15 PTFs are shown in Fig. [21](#).
188 To compare the results of PTFs 1-5 of the RF and NLR methods with those of the Rosetta
189 models, the parameters of the van Genuchten model (θ_r , θ_s , α and n) were estimated by the PTFs
190 built in the Rosetta software (PTFs 1-5), using the measured values of input variables based on
191 PTFs 1-5 as predictors in the Rosetta program. The estimated coefficients ~~of the~~ of the van
192 Genuchten model were used to calculate the estimated water content at matric suctions from 0 to
193 1500 kPa (estimated SWRCs). Then curve-by-curve comparison of the measured and estimated
194 SWRCs was performed with different evaluation statistics. Since there is no training step in the
195 Rosetta software, the results of the Rosetta model was only compared with the results of the
196 testing step. To evaluate the effect of using different descriptors of PSD on the prediction of the
197 SWRC, PTFs 6, 7, 8 and 9 from the second step were compared with PTFs 2, 3, 4 and 5 from the
198 first step, respectively (Fig. [21](#)). In the same way, to evaluate effect of using different descriptors
199 of soil structure on the prediction of the SWRC, PTFs 10, 11 and 12 from the third step were
200 compared with PTFs 3, 4 and -5 from the first step, respectively. Also, the PTFs 13-15 were

201 compared with the PTFs 4 and 5 to find out the efficiency of OM and K_s variables as predictors
202 (Fig. 21).

203 **Fig 21.**

204
205 In the present study, the k-fold cross validation approach (Efron and Tibshirani, 1994) was
206 utilized to obtain training and testing datasets for each PTF. The number of folds (i. e., k) ~~h~~was
207 ~~been~~obtained by trial and error. To do so, some PTFs, ~~which were~~selected randomly, ~~have~~
208 ~~been~~were developed with 10, 15 and 20-fold cross-validation. Then, the k value which ~~was~~
209 ~~resulted in the best performance of the PTFs, ~~was~~ selected to develop all PTFs in this study. The~~
210 ~~results showed that 20-fold cross validation performed better than the other folds, in most of the~~
211 ~~PTFs (Table 1). Therefore, 20-fold cross validation was selected to develop PTFs in this study.~~
212 Based on this approach, the 223 samples were randomly divided into 20 subsets and 20 models
213 were developed by each predicting technique for each PTF. In each model, training and testing
214 datasets were based on a ratio of 19:1. Finally, the average of the results of 20 models was
215 calculated for each PTF. Therefore, all data were used for the training and testing steps of the
216 PTFs.

217 **Table 1-**

218 *2.5 Description of modeling techniques*

219 *2.5.1 Multiple nonlinear regression*

220 A ~~nonlinear regression~~NLR model based on a second-order polynomial for the prediction of the
221 response variable y from a number of $n-p$ predictors can be written as (Rawls and Brakensiek,
222 1985):

$$y = a + \sum_{i=1}^p (b_i x_i + c_i x_i^2) \quad (2)$$

223 where a is the intercept, and two regression coefficients b_i and c_i are determined for every input
 224 variable x_i .

225

226 2.5.2 *Random forest: an ensemble of regression trees*

227 RF has become a popular tool for regression and classification problems. The RF is an ensemble
 228 method based on the regression tree methodology (i.e., ~~classification and regression trees~~
 229 ~~(CART)~~) that was introduced for better performance (Breiman, 2001). The model building
 230 process in the RF is the same as that in the CART method but without pruning (Breiman, 1984).

231 Also, whereas a regression tree only grows by a single tree, ~~but~~ the RF grows by forest of trees.

232 In other words, unlike a regression tree, in the RF for each tree only a subset of the input
 233 variables is applied. The number of inputs in each tree and also the number of trees in the forest

234 can be distinct and it depends on the dataset. Least-squares boosting (LSBoost) fits regression
 235 ensembles. At every step, the ensemble fits a new learner to the difference between the observed
 236 response and the aggregated prediction of all learners grown previously. The ensemble fits to

237 minimize the mean-squared error (MathWorks, 2018). The number of trees used here was 16

238 which was established by trial and error. An architecture of the RF algorithm is shown in Fig. 3-2

239 where input matrix X consists of N samples and M input variables (sample set $S = [(x_i, y_i), i = 1,$

240 $2, \dots, N], (X, Y) \in \mathbb{R}^{M \times R}$). The bootstrap method is utilized to construct n tree sample sets

241 from the sample set S . At each bootstrap sample, about one-third of the dataset S was utilized as

242 out of the bootstrap data or out-of-bag (*OOB*) data; whereas the rest is called in-bag data

243 (Ibrahim and Khatib, 2017) (Fig. 32). Modeling of the regression tree is done for each sample

244 set. In the RF algorithm, all individual trees give a predictive result. The final prediction value is

245 calculated based on an average result of all individual trees (Wiesmeier et al., 2011). The
246 prediction error is defined as follows (Liaw and Wiener, 2002):

$$MSE_{OOB} = \frac{\sum_{i=1}^{n_{tree}} (y_i - \hat{y}_i^{OOB})^2}{n_{tree}} \quad (3)$$

247 where MSE_{OOB} is the mean square error of the *OOB* data prediction, n_{tree} is the number of trees,
248 and y_i and \hat{y}_i^{OOB} are the actual value of the *OOB* data and the average of all *OOB* predictions,
249 respectively. Among all the ensemble methods, the RF method has high capability in solving
250 classification and regression problems, because the RF method combines several simple
251 regression trees to better optimize prediction (Zaklouta and Stanculescu, 2012). The RF method
252 increases differences for each single tree through random selection of the training samples and
253 different variables at each splitting node. In the present study, the NLR and RF algorithms were
254 implemented by `fitnlm` and `fitensemble` functions in the MATLAB software, respectively.
255 (MathWorks, 2018).

256 **Fig. 32.**
257

258 2.6 Evaluation criteria

259 The estimated water content was computed by estimated parameters of the van Genuchten model
260 for each PTF at matric suctions from 0 to 1500 kPa. For curve-by-curve comparison of the
261 measured and predicted SWRCs, different evaluation statistics were used. Various statistical
262 criteria including integral root mean square error (*IRMSE*), integral mean error (*IME*) (Tietje and
263 Tapkenhinrichs, 1993), Akaike's information criterion (*AIC*) (Akaike, 1974) and coefficient of
264 determination (R^2) (Wösten et al., 2001), were utilized to assess the predictive ability of the RF
265 and NLR algorithms, which are defined as:

$$IRMSE (cm^3 cm^{-3}) = \left[\frac{1}{b-a} \int_a^b (\hat{y}_i - y_i)^2 d \log|h| \right]^{\frac{1}{2}} \quad (4)$$

$$IME (cm^3 cm^{-3}) = \frac{1}{b-a} \int_a^b (\hat{y}_i - y_i) d \log|h| \quad (5)$$

$$AIC = N \times \ln \left[\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N} \right] + 2P \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2} \quad (7)$$

266

267 where h is the matric suction (kPa), y_i , \hat{y}_i and \bar{y}_i are the measured, predicted and average of
 268 the measured values of the water content, respectively, a and b values define the matric suction
 269 range over which the experimental curve is measured, i.e., 0 and 1500 kPa, respectively, and P
 270 and N are the number of parameters and the number of points that were considered in the SWRC,
 271 respectively. In calculating the AIC , N is the total number of points that were considered in the
 272 SWRC of all soil samples (i. e., N = number of soil samples \times number of paired points of the
 273 suction-water content for each soil sample), and i is paired points of the suctions-water content
 274 of the SWRC of each soil sample.

275 To evaluate the performance of each method in different PTFs, the effect of methods as the first
 276 factor at two levels; in the training step (i.e., NLR and RF methods) in the training step and at
 277 three levels; in the testing step (i.e., NLR, RF and Rosetta methods) in the testing step, and the
 278 different PTFs as the second factor at 15 levels (PTF1 to PTF15), were investigated using a two-
 279 way analysis of variance (ANOVA) with a randomized complete block design as a factorial test.

280 based on the *IRMSE* of prediction of the SWRC. ~~On the other hand, the~~ *IRMSE* criterion
281 calculates the total error, including bias and random errors, and is a more appropriate criterion
282 for evaluating the accuracy and reliability of the RF and NLR methods compared to other criteria
283 (Chai and Draxler, 2014). Therefore, ~~to~~ To compare the predicting accuracy and reliability of the
284 RF and NLR methods, the average values of the *IRMSE* was compared with Duncan's test by
285 MathWorks (2018) software.

286

287 **3 Results and discussion**

288 *3.1 Descriptive statistics of the soil properties*

289 Table 4-2 summarizes some basic descriptive statistics for soil variables of the entire dataset used
290 for the development of the PTFs. It can be seen that the average and maximum of clay content
291 were 21.4 and 48%, respectively. ~~It can be seen that the average clay content was 21.4 %, and~~
292 ~~exceeded 50%.~~ The OM ranged from 0.17 to 4.41% with a mean of 1.84%, which ~~was~~ low due
293 to the arid and semi-arid climates of Iran. The variation ~~of the~~ soil texture is shown graphically
294 in the United States Department of Agriculture (USDA) textural triangle (Fig. 4-3). Considering
295 the distribution and range of the variables (Fig. 4-3 and Table 4-2), the dataset can be considered
296 as representative of soils in arid and semi-arid regions of Iran.

297

Table 4-2

298

Fig. 4-3.

299 *3.2 Correlation of input and output variables*

300 The simple correlation coefficients between all variables are depicted by matrix plot in Fig. 4-4.
301 Correlation analysis was done between normalized input and output variables. The correlation test
302 was not performed for the θ_r variable, because its value was zero in 138 out of 223 soil samples.

303 ~~Also the zero value as hasve~~ been reported in ~~some~~ other studies (Campbell and Horton Jr, 2002;
304 Rawls et al., 1991; Tomasella et al., 2000) for θ_r variable. Clay and sand contents, θ_{FC} , θ_{PWP} , d_g
305 and OM had the greatest significant correlations with the parameters of the van Genuchten model
306 (Fig. 14), which ~~are~~ was consistent with other studies (Dexter et al., 2008; Nemes et al., 2006). For
307 example, the correlation coefficients between clay content and θ_s ($r = 0.323$) is close to that
308 between ~~the~~ OM and θ_s ($r = 0.268$). Also, the results showed that there were significant correlations
309 between θ_{PWP} and input variables of clay content (+), sand content (—), BD (—), OM (+) and K_s
310 (—), and also between θ_{PWP} and θ_s (+) and n (—) parameters of the van Genuchten model (Fig. 4).
311 Botula et al. (2012) also found the same observation for the correlation of θ_{PWP} with sand and clay
312 contents and BD of tropical Lower Congo soils. Nevertheless, with regard to these correlation
313 coefficients, clay and sand contents, θ_{FC} , d_g and OM can be used for developing PTFs to estimate
314 the SWRC. On the contrary, there was no correlation between K_s and the van Genuchten model
315 parameters. There are many cases, where two variables might not show a strong simple correlation,
316 but may show a strong association in the regression, along with other predictors. In other words,
317 the simple correlation coefficient is a way to show the relationship between ~~two~~ independent and
318 dependent variables, but it cannot show a model for the relationship between these two variables,
319 when other independent variables have been used in a multiple regression (Simmons et al., 2011).
320 The results of multiple regression analysis with backward selection method showed that the K_s
321 variable remained in the PTF14 and PTF15 for all the van Genuchten model parameters. Some of
322 the regression equations with backward selection method are shown in the following as examples:

$$\theta_r = -0.69 + 0.22 \times \text{Clay} + 0.278 \times \text{Sand} + 0.20 \times K_s, R = 0.31^{**} \quad (8)$$

$$\alpha = -3.72 + 0.23 \times \text{Clay} + 0.17 \times \text{BD} + 0.282 \times K_s, R = 0.33^{**} \quad (9)$$

$$n=-1.76+0.24\times\text{Sand}+0.164\times K_s, R=0.30^{**} \quad (10)$$

On the other hand, the non-linear correlations between variables are very important in this study. Because, both the multiple nonlinear regression NLR approach and random forest RF data mining technique, which were used, are non-linear prediction methods. Fig. 4 only shows simple linear correlation between variables, but there may be non-linear correlations between variables, which may affect the estimation of the dependent variables. For example, the results of non-linear correlations showed that K_s had strong correlations with θ_s and α of the van Genuchten model parameters by logarithmic ($\theta_s=0.652-0.027\times\ln K_s, R=0.62^{**}$) and power ($\alpha=0.007\times K_s^{0.283}, R=0.57^{**}$) equations, respectively, and these non-linear correlations which were increased mostly in comparison with greater than their simple correlations, indicating nonlinear relationships of the K_s with θ_s and α . Therefore, regression method can discover and apply the law that exists between these two variables.

Fig. 41.

3.3 Development of the PTFs using the RF and NLR methods

Results of the multicollinearity analysis (VIF) are shown in Table 23. The VIF values in Table 2 showed low levels of multicollinearity among the independent variables ($VIF<10$) (Khodaverdiloo et al., 2011).

Table 23-

344 3.3.1 Comparing the accuracy and reliability of the RF and NLR methods

345 ~~Table 34 shows the results of analysis of variance the ANOVA of the IRMSE of prediction of the~~
346 ~~SWRC by different methods and PTFs. The analysis of variance showed that the effect of type of~~
347 ~~methods and PTFs, and their interaction, on the IRMSE was significant at $P < 0.01$, 0.01 and~~
348 ~~0.05, respectively, in the training step, and also at $P < 0.01$, 0.01 and 0.01, respectively, in the~~
349 ~~testing step. Therefore, we focus on the results and discussion of the mean comparison was~~
350 ~~performed and results and discussion were written according to of the method \times PTF interaction~~
351 ~~effects.~~

352 **Table 34-**

353 Results of the prediction of the SWRC through the van Genuchten model using the NLR and RF-
354 based PTFs are depicted in Figs. 5 and 6 for the training and testing steps, respectively. The
355 accuracy and reliability are used to express the performance of the PTFs in the
356 training and testing steps, respectively.

357 The results of the first to fourth steps of the training dataset (Fig. 5) showed that the RF method
358 had better performance compared to the NLR method for the prediction of the SWRC in all PTFs
359 in terms of the *IRMSE* and R^2 criteria and the differences were significant ($P < 0.05$) for PTFs 2,
360 3, 6, 7, 10, 13, 14 and 15 in terms of the *IRMSE* criterion. Also, the accuracy of the RF method
361 was better than that of the NLR method in 80% of the PTFs (with the exception of the PTFs 5, 9
362 and 12) in terms of the *AIC* criterion. In the training step, the values of the *IRMSE* of the first to
363 fourth steps for the NLR model varied from 0.030 to 0.063 $\text{cm}^3 \cdot \text{cm}^{-3}$ and these were larger than
364 those in the RF model, which ranged from 0.028 to 0.061 $\text{cm}^3 \cdot \text{cm}^{-3}$, respectively. Also, the
365 values of the R^2 of the first to fourth steps for the RF model varied from 0.981 to 0.992, and this
366 was larger than those in the NLR model, which ranged from 0.979 to 0.991 (Fig. 5).

367 The results of the first to fourth steps of the testing dataset (Fig. 6) showed that the NLR method
368 had a better performance compared to the RF method on the prediction of the SWRC for PTFs 5,
369 8, 9 and 15 only in terms of the *IRMSE* criterion (significant at $P < 0.05$). In the other PTFs
370 there were no significant differences between the *IRMSE* of the two methods and the R^2 and *AIC*
371 criteria were comparable. In the testing step, the values of the *IRMSE* and *AIC* of the first to
372 fourth steps for the RF models varied from 0.038 to 0.065 $\text{cm}^3 \text{cm}^{-3}$ and from -13476.2 to -
373 17646.8, respectively, and these were comparable to those of the NLR models (with the
374 exception of the PTF1), which ranged from 0.032 to 0.064 $\text{cm}^3 \text{cm}^{-3}$ and from -14096.1 to -
375 19234.1, respectively (Fig. 6). Also, the values of the R^2 of the first to fourth steps for the NLR
376 models varied from 0.979 to 0.989, and this was comparable to those of the RF models for all
377 PTFs, which ranged from 0.977 to 0.987 (Fig. 6).

378 In each of the PTFs 1 to 5, the NLR and RF methods performed better ($P < 0.05$) than the Rosetta
379 PTFs. Fig. 6(A) shows that the Rosetta-based PTFs have had greater values of the *IME* criterion
380 compared to the NLR and RF-based PTFs. The reason can be attributed to the various methods
381 of optimizing parameters. The Rosetta method has only one artificial neural network (ANN) type
382 with particular structure. In other words, the number of hidden layers (one) and neurons (six) and
383 also the activation function (tangent hyperbolic) are constant for prediction of the SWRC in the
384 Rosetta software. Therefore, the Rosetta method is not a dynamic approach for optimization,
385 whereas the parameters of the RF method, such as number of splits and trees, and learning rate
386 continuously and dynamically, change to achieve the best result of the objective function. The
387 Rosetta method was developed from a quite-large dataset, while the soils used in the present
388 study were collected from a completely different climate area that was not represented in the
389 Rosetta's database. Also, presented RF and NLR models were trained using this particular dataset

390 while Rosetta had been trained using a different dataset. In other words, the results of the PTFs
391 in the testing step ~~are-were~~ based on a soil dataset used for training. This could be a reason for
392 Rosetta's poor performance compared with the RF and NLR methods. As a result, it seems that
393 the universal portability of the Rosetta method can be limited. The testing results are in
394 agreement with Touil et al. (2016) who found that the parametric-based PTFs of nonlinear
395 models, gave a better prediction than the Rosetta PTFs. The Figs. 5(A) and 6(A) showed that all
396 of the *IME* values were negative for all PTFs at the training and testing steps. There are regular
397 errors (bias) in the prediction of the SWRC that can be corrected by finding a correction
398 coefficient, which would improve the accuracy and reliability of the estimations (Bayat et al.,
399 2015).

400 **Fig. 5.**

401 **Fig. 6.**

402
403 The RF method in the training section gave better predictions of the SWRC compared to the
404 NLR method ~~for the prediction of the SWRC~~ (Fig. 5). The RF method produces low-bias and
405 variation ~~results~~ in the data by majority voting compared to a single regression tree (Cheng et al.,
406 2019; Matin and Chelgani, 2016). In this connection, the results of the standard deviations (SD)
407 of evaluation criteria in each PTF for the training step (Fig. 5) showed that the RF method had a
408 lower SD-variation than the NLR method. Accordingly, the values of SD for the *IRMSE* and R^2
409 criteria were 0.024 and 0.022, respectively, for the NLR model, and these were larger than those
410 in the RF model, which were 0.020 and 0.017, respectively, for the training step. On the other
411 hand, the RF method can be applied to high dimensional datasets in regressions (Janitza et al.,
412 2016; Zhao et al., 2016).

413 As depicted in Fig. 6, unlike ~~in~~ the training section, the NLR method gave better predictions in
414 the testing section, compared to the RF method for the prediction of the SWRC. In other words,
415 the reliability of the NLR method was better than that of the RF method in all the PTFs. The
416 ~~nonlinear regression~~NLR equations can be more useful than the MLR method for the prediction
417 of the SWRC due to their high flexibility (Williams et al., 1992). In other words, the NLR
418 models have capacity to capture nonlinear relationships in the dataset. Tomasella et al. (2000)
419 successfully developed parametric-PTFs for soils of the humid tropics using polynomials of n^{th}
420 order. Medrado and Lima (2014) successfully developed NLR-based PTFs to predict the four
421 parameters of the van Genuchten model for Brazilian soils. Also, Touil et al. (2016) developed
422 parametric-PTFs to predict the SWRC using the NLR method from more readily-available
423 properties such as soil texture, OM content, and BD for 242 soil samples of Algeria. They
424 reported that the parametric-PTFs had better performance ~~compared to the~~than Rosetta-based
425 PTFs.

426 In the present study, in contrast to the NLR method, which had less differences between the error
427 values of the training and testing steps, ~~the, the~~ error values of the RF method in the testing
428 dataset ~~was-were~~ much greater than those in the training dataset. These results can be due to
429 overprediction phenomenon in the RF method. Gupta et al. (2017) expressed that one of the
430 disadvantages of the RF method is the overprediction. In other words, the RF method is a
431 ‘greedy’ method that easily leads to overprediction and instability in the testing step and solving
432 this problem can be of great significance for improving the reliability of the RF method (Liu,
433 2014). Also, Ma et al. (2005) reported ~~the~~ instability in results of the RF method. The forest size
434 developed by the RF has not been clearly defined (Liu, 2014). Therefore, oversized scale can
435 decrease the reliability and efficiency of the SWRC prediction. Hong et al. (2016) evaluated

436 landslide susceptibility maps produced using the RF method and compared these maps with
437 those from statistical-based methods, such as logistic regression, and their study revealed that the
438 performance of the statistical-based methods was better than that of the RF method. ~~Also, a~~
439 similar result was reported by Esposito et al. (2014). Generally, RFs are best suited for problems
440 with many input variables and a reasonable sample size. According to ~~the our~~ results (Figs. 5 and
441 6), performance of the PTFs was improved by increasing the number of input variables.

442 3.3.2 Evaluation of the effect of the basic soil properties on prediction performance of the 443 SWRC

444 A significant improvement was achieved in the accuracy of PTF5 (with the inputs of Sand
445 content+Clay content+BD+ θ_{FC} + θ_{PWP}) compared to other PTFs (with the exception of PTFs 4, 8,
446 9, 11 and 12) by both NLR and RF methods in terms of the *IRMSE* criterion (Fig. 5). Among the
447 PTFs of each method (RF or NLR), PTF5 had the greatest R^2 (0.992 and 0.991, respectively) and
448 the smallest *IRMSE* (0.028 and 0.03, respectively) and *AIC* (-19432 and -19571.1, respectively)
449 ~~values~~, in the training step of the prediction of the SWRC. In connection with the importance of
450 input variables, an improvement was achieved in the reliability of the prediction of the SWRC by
451 PTFs 9 (with the inputs of $d_g + \delta_g + BD + \theta_{FC} + \theta_{PWP}$) and 12 (with the inputs of Sand content+Clay
452 content+TP+ θ_{FC} + θ_{PWP}) from the second and third steps, using the NLR (*IRMSE*=0.032 cm³·cm⁻³,
453 *AIC*=-19234.1 and R^2 =0.989) and RF (*IRMSE*=0.038 cm³·cm⁻³, *AIC*=-17646.8 and R^2 =0.987)
454 methods, respectively, in comparison with the other PTFs of each method (Fig 6). However, the
455 differences of ~~the~~ PTFs 9 and 12 were not significant ($P < 0.05$) with PTFs 4, 5, 8, 11 and 12 in
456 the NLR method and ~~also~~ with PTFs 4, 5, 8, 9 and 11 in the RF method, respectively, in terms of
457 the *IRMSE* criterion.

458

459 3.3.2.1 Effect of using different input variables of PSD and soil structure as predictors on the
460 SWRC prediction

461 ~~Input variables such as textural contents (clay and sand contents) and statistics (d_g and δ_g) as~~
462 ~~different descriptors of the PSD, and also the TP and BD as different descriptors of the soil~~
463 ~~structure, were used for prediction of the SWRC. Thus, t~~To evaluate the effect of using different
464 descriptors of the PSD on the prediction of the SWRC, PTFs 2, 3, 4 and 5 (clay and sand
465 contents) from the first step were compared with PTFs 6, 7, 8 and 9 (d_g and δ_g) from the second
466 step, respectively. In the same way, to evaluate the effect of using different descriptors of ~~the~~ soil
467 structure on the prediction of the SWRC, PTFs 3, 4 and 5 (BD) were compared with PTFs 10, 11
468 and 12 (TP) from the third step, respectively. The accuracy and reliability of the prediction of the
469 SWRC by both NLR and RF methods were not significantly different ($P < 0.05$) (Figs. 5B and
470 6B). For descriptors of soil structure, the accuracy and reliability of the prediction of the SWRC
471 by both NLR and RF methods decreased in terms of the *IRMSE* criterion for PTFs 10 to 12 from
472 the third step compared to PTFs 3 to 5 (with the exception of PTFs 11 and 12 in the testing step
473 for the RF method), respectively, when TP was used instead of BD in the list of input variables
474 (Figs. 5B and 6B). However, the differences were not significant ($P < 0.05$).

475 The lack of significant differences s between textural contents (clay and sand contents) and
476 statistics (d_g and δ_g), and also between TP and BD on the SWRC prediction can be due to
477 correlation of these parameters with the parameters of the van Genuchten model (Fig. 14). The
478 SWRC ~~can be~~ is strongly influenced by the soil structure or pore-size distribution and soil texture
479 at small and great matric suctions, respectively (Pachepsky et al., 2006). Therefore, input
480 variables of the textural contents or statistics can influence the residual saturation region of the
481 SWRC. However, soil water content at the dry end (high matric suctions) of the SWRC is

482 primarily determined by textural contents (Hillel, 1998) which had significant correlations with θ_r
483 parameter (with the exception of the clay content) (Fig. 1). Also, TP and BD are indicators of
484 soil structure and had significant correlations with θ_s (Fig. 14). Indeed, TP was calculated by BD
485 and particle density (Rab et al., 2011). ~~On the other hand, the~~ d_g and δ_g predictors were derived
486 from soil textural contents (Shirazi and Boersma, 1984). ~~In other words, textural contents data~~
487 ~~can be converted to d_g and δ_g by equations of Shirazi and Boersma (1984). Also, TP was~~
488 ~~calculated by BD and particle density (Rab et al., 2011).~~ Therefore, these could be reasons for
489 similar effects of textural contents and statistics and also TP and BD predictors on the prediction
490 of the SWRC.

491 Many researchers used textural contents (Adhikary et al., 2008; Chakraborty et al., 2011;
492 Minasny et al., 1999; Tomasella and Hodnett, 1998), d_g and δ_g (Rab et al., 2011; Scheinost et al.,
493 1997; Ungaro et al., 2005), BD (Bayat et al., 2011; Pachepsky et al., 1998) and TP (Bayat et al.,
494 2011; Pachepsky et al., 1998; Schaap et al., 1998) as effective predictors to derive point- and
495 parametric-PTFs. Nemes et al. (2003), Schaap et al. (2001) and Schaap et al. (1998) reported that
496 the variables of PTF5 have better capability on predicting the parameters of the van Genuchten
497 (1980) model with an average *RMSE* of 0.026, 0.044 and 0.058 $\text{cm}^3\text{cm}^{-3}$, respectively.

498 According to the results of the accuracy (Fig. 5) and reliability (Fig. 6) of PTFs 5, 9 and 12, it
499 seems that certain points of the SWRC (e.g., θ_{FC}) can help to improve the prediction of the
500 SWRC and this is in agreement with Schaap et al. (2001). These results indicate that the presence
501 of at least one moisture points (e.g., θ_{FC}) can improve the prediction of the SWRC. In ~~other~~
502 ~~words, according to the results of the accuracy (Fig. 5) and reliability (Fig. 6) of the NLR and RF~~
503 ~~methods for different PTFs, at least one moisture point is necessary to predict the SWRC. For~~
504 ~~example, in~~ the first step, PTF5 with two moisture points ($\theta_{FC} + \theta_{PWP}$) and PTF4 with one

505 moisture point (θ_{FC}) improved the prediction of the SWRC by 55, 48, 42% and 51, 44, 38% in
506 terms of the *IRMSE* criterion compared to the PTFs 1, 2 and 3, respectively, in the RF method in
507 the training step. In the testing section of the second step, PTF9 with two moisture points
508 ($\theta_{FC}+\theta_{PWP}$) and PTF8 with one moisture point (θ_{FC}) decreased the *IRMSE* by 49, 44% and 44,
509 39% compared to PTFs 6 and 7, respectively, in the NLR method. The points above are also true
510 for the RF-based PTF12 in the third step of the testing section. Many researchers successfully
511 applied θ_{FC} and θ_{PWP} as effective predictors to derive point- and parametric-PTFs (Børgesen and
512 Schaap, 2005; Nemes et al., 2003; Schaap et al., 2001; Touil et al., 2016; Twarakavi et al., 2009).

513

514 3.3.2.2 *Effect of using OM and K_s as predictors on the SWRC prediction*

515 To evaluate the effect of using OM and/or K_s and points of the SWRC on the prediction of the
516 SWRC, the performances of PTFs 13, 14 and 15 were compared with those of PTFs 4 and 5. The
517 accuracy and reliability of the prediction of the SWRC by both NLR and RF methods,
518 significantly ($P<0.05$) decreased in terms of the *IRMSE*, for the PTFs 13, 14 and 15 from the
519 fourth step, when OM and/or K_s were used with textural contents and BD as inputs instead of θ_{FC}
520 or both θ_{FC} and θ_{PWP} in the list of input variables, compared to PTFs 4 and 5 at the first step
521 (Figs. 5B and 6B). Therefore OM and K_s were not as effective predictors as θ_{FC} and θ_{PWP} in the
522 prediction of the SWRC, because θ_{FC} and θ_{PWP} are two points of the SWRC and enter direct
523 information of the SWRC into the PTFs, whereas OM and K_s enter indirect information, and
524 therefore had less effect in the improvement of the estimation of the SWRC. These results agreed
525 well with results obtained by Børgesen and Schaap (2005). They reported that PTFs with the
526 inputs of θ_{FC} and θ_{PWP} had smaller *RMSE* values than a PTF with the input of OM (0.038 versus
527 0.042) in the prediction of the SWRC. On the other hand, the results showed that by adding OM

528 and/or K_s as predictors in the PTFs 13, 14 and 15, the accuracy (Fig. 5B) and reliability (Fig. 6B)
529 of the prediction of the SWRC improved by 16, 13, 17 and 7.1, 6.3, 6.9%, respectively,
530 compared to the PTF3 in terms of the *IRMSE* criterion in the RF method.
531 The SWRC depends mainly on the soil texture and structure (Hillel, 1998), with OM affecting
532 the SWRC through development of soil structure (Nemes et al., 2005), important at low suctions.
533 However, the OM retains water itself. Similarly, K_s can be a descriptive index of soil texture and
534 porosity (Hillel, 1998). The correlation results showed (Fig. 14) that K_s can be strongly
535 influenced by clay content and textural statistics (d_g and δ_g) soil texture and TP(Fig. 4). Bayat et
536 al. (2013b) applied OM and K_s to estimate water content at the measured matric suctions. They
537 found that the OM and K_s can be most appropriately used in point-based PTFs to estimate water
538 content at the matric suctions of 25 and 50 kPa. Also, the result of the present study agreed well
539 ~~to the~~with results obtained by Hollis et al. (1977) and Rawls et al. (1983). In this study, the OM
540 and K_s in the PTFs 13, 14 and 15 were not effective predictors compared to θ_{FC} and θ_{PWP} in the
541 PTFs 4 and 5, otherwise they had better results than PTF3.

542

543 **4 Conclusion**

544 Machine-learning tools have been widely applied for the prediction of the SWRC. The present
545 study evaluated the capability and performance of the RF method as a novel machine learning
546 tool and compared its performance with that of the ~~nonlinear regression~~-(NLR) method on the
547 prediction of the SWRC, using different combinations of easily-available soil properties. It was
548 found that the RF method had a better performance ($P<0.05$) than the NLR method in the
549 training step of the prediction of the SWRC in term of the *IRMSE*, *AIC* and R^2 criteria. However,
550 in the testing step, NLR had a better performance than RF. The poor performance of the RF

551 compared to the NLR method could be due to overprediction in the former, resulting in
552 instability in the testing step. The RF method can be sensitive to sparse areas on the prediction
553 space. In other words, the performance and, sensitivity of predictions, and the computational
554 intensity of the RF method depends on the distribution and number of observations and input
555 variables. Therefore, ~~this~~ the method should be tested further with different datasets to evaluate
556 its performance through soil and water investigations. An improvement was achieved in the
557 accuracy of the prediction of the SWRC in the training step of the PTF5 (with the inputs of Sand
558 content+Clay content+BD+ θ_{FC} + θ_{PWP}) by both NLR and RF methods and also an improvement
559 was achieved in the reliability of the PTF9 (with the inputs of d_g + δ_g +BD+ θ_{FC} + θ_{PWP}) and PTF12
560 (with the inputs of Sand content +Clay content+TP+ θ_{FC} + θ_{PWP}) by the NLR and RF methods
561 compared to other PTFs, respectively. Considering that the PTFs 5, 9, and 12 had no significant
562 difference from PTF4 (with the inputs of Sand content+Clay content+BD+ θ_{FC}) and PTF8 (with
563 the inputs of d_g + δ_g +BD+ θ_{FC} + θ_{PWP}), these latter PTFs, with less and more-easily measured input
564 variables, are suggested to be the best PTFs for the prediction of the SWRC. Also, PTFs without
565 predictors of θ_{FC} and θ_{PWP} , such as the PTF3 (with the inputs of Sand content+Clay content+BD)
566 and PTF7 (with the inputs of d_g + δ_g +BD), can be effective models for the prediction of the
567 SWRC.

568

569 **Acknowledgements**

570 This work was funded by Bu Ali Sina University, Hamedan, Iran. The authors are deeply
571 grateful to anonymous reviewers and the editor for their helpful comments on the manuscript.

572

573

574

575

576 **References**

- 577 Adhikary, P.P., Chakraborty, D., Kalra, N., Sachdev, C., Patra, A., Kumar, S., Tomar, R.,
578 Chandna, P., Raghav, D., Agrawal, K., 2008. Pedotransfer functions for predicting the
579 hydraulic properties of Indian soils. *Soil Res.* 46, 476-484.
- 580 Akaike, H., 1974. A new look at the statistical model identification. *IEEE transactions on*
581 *automatic control* 19, 716-723.
- 582 Araya, S.N., Ghezzehei, T.A., 2019. Using Machine Learning for Prediction of Saturated
583 Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. *Water Resour.*
584 *Res.* 55, 5715-5737.
- 585 Bayat, H., Ersahin, S., Hepper, E.N., 2013a. Improving estimation of specific surface area by
586 artificial neural network ensembles using fractal and particle size distribution curve
587 parameters as predictors. *Environ. Model Assess.* 18, 605-614.
- 588 Bayat, H., Neyshabouri, M., Mohammadi, K., Nariman-Zadeh, N., 2011. Estimating water
589 retention with pedotransfer functions using multi-objective group method of data
590 handling and ANNs. *Pedosphere* 21, 107-114.
- 591 Bayat, H., Neyshaburi, M.R., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2013b.
592 Improving water content estimations using penetration resistance and principal
593 component analysis. *Soil Tillage Res.* 129, 83-92.
- 594 Bayat, H., Sedaghat, A., Sinegani, A.A.S., Gregory, A.S., 2015. Investigating the relationship
595 between unsaturated hydraulic conductivity curve and confined compression curve. *J.*
596 *Hydrol.* 522, 353-368.

597 Berry, W.D., 1993. Understanding regression assumptions. Sage Publications, London.

598 Blake, G., Hartge, K., 1986. Bulk density, Methods of Soil Analysis: Part 1. Physical and
599 Mineralogical Methods, Madison, Wisconsin, USA: Soil Sci. Soc. Am. J.

600 Børgesen, C.D., Schaap, M.G., 2005. Point and parameter pedotransfer functions for water
601 retention predictions for Danish soils. *Geoderma* 127, 154-167.

602 Botula, Y.-D., Cornelis, W., Baert, G., Van Ranst, E., 2012. Evaluation of pedotransfer functions
603 for predicting water retention of soils in Lower Congo (DR Congo). *Agric. Water Manag.*
604 111, 1-10.

605 Botula, Y.-D., Cornelis, W.M., Baert, G., Mafuka, P., Van Ranst, E., 2013. Particle size
606 distribution models for soils of the humid tropics. *Journal of Soils and Sediments* 13,
607 686-698.

608 Bouma, J., 1989. Using soil survey data for quantitative land evaluation, *Advances in soil
609 science*. Springer, pp. 177-213.

610 Breiman, L., 1984. Classification and regression trees. Routledge, New York.

611 Breiman, L., 2001. Random forests. *Machine learning* 45, 5-32.

612 Bruce, R.R., Luxmoore, R.J., 1986. Water Retention: Field Methods, In: Klute, A. (Ed.),
613 Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods. Soil Science
614 Society of America, American Society of Agronomy, Madison, WI, pp. 663-686.

615 Campbell, G.S., Horton Jr, R., 2002. Methods of Soil Analysis: Part 4, Physical Methods. Soil
616 Sci. Soc. Am.

617 Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?—
618 Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7, 1247-1250.

619 Chakraborty, D., Mazumdar, S., Garg, R., Banerjee, S., Santra, P., Singh, R., Tomar, R., 2011.
620 Pedotransfer functions for predicting points on the moisture retention curve of Indian
621 soils. *Indian J. Agr. Sci.* 81, 1030.

622 Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method
623 approach to model travel mode choice behavior. *Travel behaviour and society* 14, 1-10.

624 Dexter, A., Czyż, E., Richard, G., Reszkowska, A., 2008. A user-friendly water retention
625 function that takes account of the textural and structural pore spaces in soil. *Geoderma*
626 143, 243-253.

627 Dobarco, M.R., Cousin, I., Le Bas, C., Martin, M.P., 2019. Pedotransfer functions for predicting
628 available water capacity in French soils, their applicability domain and associated
629 uncertainty. *Geoderma* 336, 81-95.

630 Efron, B., Tibshirani, R.J., 1994. *An introduction to the bootstrap*. CRC press.

631 Esposito, C., Barra, A., Evans, S.G., Scarascia Mugnozza, G., Delaney, K., 2014. Landslide
632 susceptibility analysis by the comparison and integration of random forest and logistic
633 regression methods; application to the disaster of Nova Friburgo-Rio de Janeiro, Brasil
634 (January 2011), EGU General Assembly Conference Abstracts.

635 Gee, G.W., Or, D., 2002. 2.4 Particle-Size Analysis, In: Dane, J.H., Topp, C.G. (Eds.), *Methods*
636 *of Soil Analysis: Part 4 Physical Methods*. Soil Science Society of America, Madison,
637 WI, pp. 255-293.

638 Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., 2019a. Machine Learning
639 Approaches to Develop Pedotransfer Functions for Tropical Sri Lankan Soils. *Water* 11,
640 1940.

641 Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., Amarasekara, M., 2019b.
642 Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. *Soil*
643 *Till. Res.* 190, 109-119.

644 Gupta, B., Rawat, A., Jain, A., Arora, A., Dhani, N., 2017. Analysis of various decision tree
645 algorithms for classification in data mining. *Int. J. Comput. Appl.* 163, 15-19.

646 Haghverdi, A., Leib, B.G., Cornelis, W.M., 2015. A simple nearest-neighbor technique to predict
647 the soil water retention curve. *Transactions of the ASABE* 58, 697-705.

648 Hillel, D., 1998. *Environmental soil physics: Fundamentals, applications, and environmental*
649 *considerations.* Academic press.

650 Hocking, R.R., 2013. *Methods and applications of linear models: regression and the analysis of*
651 *variance.* John Wiley & Sons.

652 Hollis, J., Jones, R., Palmer, R., 1977. The effects of organic matter and particle size on the
653 water-retention properties of some soils in the West Midlands of England. *Geoderma* 17,
654 225-238.

655 Hong, H., Pourghasemi, H.R., Pourtaghi, Z.S., 2016. Landslide susceptibility assessment in
656 Lianhua County (China): a comparison between a random forest data mining technique
657 and bivariate and multivariate statistical models. *Geomorphology* 259, 105-118.

658 IBM, C., 2016. *IBM SPSS Statistics for Windows, Version 24.0.* Armonk, NY: IBM Corp.

659 Ibrahim, I.A., Khatib, T., 2017. A novel hybrid model for hourly global solar radiation prediction
660 using random forests technique and firefly algorithm. *Energy Convers. Manag.* 138, 413-
661 425.

662 Janitza, S., Tutz, G., Boulesteix, A.-L., 2016. Random forest for ordinal responses: prediction
663 and variable selection. *Comput. Statist. Data Anal.* 96, 57-73.

664 Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced
665 pedotransfer functions with support vector machines to predict water retention of
666 calcareous soil. *Eur. J. Soil Sci.* 67, 276-284.

667 Khodaverdiloo, H., Homaei, M., van Genuchten, M.T., Dashtaki, S.G., 2011. Deriving and
668 validating pedotransfer functions for some calcareous soils. *J. Hydrol.* 399, 93-99.

669 Klute, A., 1986. Water Retention: Laboratory Methods, In: Klute, A. (Ed.), *Methods of Soil*
670 *Analysis: Part 1—Physical and Mineralogical Methods.* Soil Science Society of America,
671 American Society of Agronomy, Madison, WI, pp. 635-662.

672 Klute, A., Dirksen, C., 1986. Hydraulic Conductivity and Diffusivity: Laboratory Methods, In:
673 Klute, A. (Ed.), *Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods.*
674 Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 687-
675 734.

676 Koekkoek, E., Booltink, H., 1999. Neural network models to predict soil water retention. *Eur. J.*
677 *Soil Sci.* 50, 489-495.

678 Lamorski, K., Pachepsky, Y., Sławiński, C., Walczak, R., 2008. Using support vector machines
679 to develop pedotransfer functions for water retention of soils in Poland. *Soil Sci. Soc.*
680 *Am. J.* 72, 1243-1247.

681 Lamorski, K., Sławiński, C., Moreno, F., Barna, G., Skierucha, W., Arrue, J.L., 2014. Modelling
682 soil water retention using support vector machines with genetic algorithm optimisation.
683 *Sci. World J.* 2014, 740521, 1-10.

684 Liaw, A., Wiener, M., 2002. Classification and regression by random forest. *R news* 2, 18-22.

685 Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture:
686 comparison of regression tree and Random Forest models. *Geoderma* 170, 70-79.

687 Liu, Y., 2014. Random forest algorithm in big data environment. *Comput. Model. New Tech.* 18,
688 147-151.

689 Ma, Y., Cukic, B., Singh, H., 2005. A classification approach to multi-biometric score fusion,
690 *International Conference on Audio-and Video-Based Biometric Person Authentication.*
691 Springer, pp. 484-493.

692 MathWorks, 2018. MATLAB: the language of technical computing, Inc., Natick, Massachusetts,
693 United States.

694 Matin, S., Chelgani, S.C., 2016. Estimation of coal gross calorific value based on various
695 analyses by random forest method. *Fuel* 177, 274-278.

696 Medrado, E., Lima, J.E., 2014. Development of pedotransfer functions for estimating water
697 retention curve for tropical soils of the Brazilian savanna. *Geoderma Regional* 1, 59-66.

698 Merdun, H., Çınar, Ö., Meral, R., Apan, M., 2006. Comparison of artificial neural network and
699 regression pedotransfer functions for prediction of soil water retention and saturated
700 hydraulic conductivity. *Soil Tillage Res.* 90, 108-116.

701 Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the
702 development of pedotransfer functions for water-retention curves. *Geoderma* 93, 225-
703 253.

704 Mualem, Y., 1976. A new model for predicting the hydraulic conductivity of unsaturated porous
705 media. *Water Resour. Res.* 12, 513-522.

706 Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2005. Influence of organic matter on the estimation of
707 saturated hydraulic conductivity. *Soil Sci. Soc. Am. J.* 69, 1330-1337.

708 Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor
709 approach to estimate soil hydraulic properties. *Soil Sci. Soc. Am. J.* 70, 327-336.

710 Nemes, A., Schaap, M., Wösten, J., 2003. Functional evaluation of pedotransfer functions
711 derived from different scales of data collection. *Soil Sci. Soc. Am. J.* 67, 1093-1102.

712 Neyshaburi, M.R., Bayat, H., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2015.
713 Improvement in estimation of soil water retention using fractal parameters and
714 multiobjective group method of data handling. *Arch. Agron. Soil Sci.* 61, 257-273.

715 Nguyen, P.M., Haghverdi, A., De Pue, J., Botula, Y.-D., Le, K.V., Waegeman, W., Cornelis,
716 W.M., 2017. Comparison of statistical regression and data-mining techniques in
717 estimating soil water retention of tropical delta soils. *Biosyst. Eng.* 153, 12-27.

718 Pachepsky, Y., Rawls, W., Gimenez, D., Watt, J., 1998. Use of soil penetration resistance and
719 group method of data handling to improve soil water retention estimates. *Soil Tillage*
720 *Res.* 49, 117-126.

721 Pachepsky, Y.A., Rawls, W., 1999. Accuracy and reliability of pedotransfer functions as affected
722 by grouping soils. *Soil Sci. Soc. Am. J.* 63, 1748-1757.

723 Pachepsky, Y.A., Rawls, W., Lin, H., 2006. *Hydropedology and pedotransfer functions.*
724 *Geoderma* 131, 308-316.

725 Pachepsky, Y.A., Timlin, D., Varallyay, G., 1996. Artificial neural networks to estimate soil
726 water retention from easily measurable data. *Soil Sci. Soc. Am. J.* 60, 727-733.

727 Rab, M., Chandra, S., Fisher, P., Robinson, N., Kitching, M., Aumann, C., Imhof, M., 2011.
728 Modelling and prediction of soil water contents at field capacity and permanent wilting
729 point of dryland cropping soils. *Soil Res.* 49, 389-407.

730 Rajkai, K., Kabos, S., Van Genuchten, M.T., 2004. Estimating the water retention curve from
731 soil properties: comparison of linear, nonlinear and concomitant variable methods. *Soil*
732 *Tillage Res.* 79, 145-152.

733 Rawls, W., Brakensiek, D., Soni, B., 1983. Agricultural management effects on soil water
734 processes part I: Soil water retention and Green and Ampt infiltration parameters.
735 Transactions of the ASAE 26, 1747-1752.

736 Rawls, W., Gish, T., Brakensiek, D., 1991. Estimating soil water retention from soil physical
737 properties and characteristics, Advances in soil science. Springer, pp. 213-234.

738 Rawls, W.J., Brakensiek, D., 1985. Prediction of soil water properties for hydrologic modeling,
739 Watershed management in the eighties. ASCE, pp. 293-299.

740 Schaap, M.G., Leij, F.J., van Genuchten, M.T., 1998. Neural network analysis for hierarchical
741 prediction of soil hydraulic properties. Soil Sci. Soc. Am. J. 62, 847-855.

742 Schaap, M.G., Leij, F.J., van Genuchten, M.T., 2001. Rosetta: A computer program for
743 estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol.
744 251, 163-176.

745 Scheinost, A., Sinowski, W., Auerswald, K., 1997. Regionalization of soil water retention curves
746 in a highly variable soilscape, I. Developing a new pedotransfer function. Geoderma 78,
747 129-143.

748 Seo, S., 2006. A review and comparison of methods for detecting outliers in univariate data sets,
749 Thesis for Master of Science in Field of Public Health University of Pittsburgh, pp. 1-59.

750 Shirazi, M.A., Boersma, L., 1984. A unifying quantitative analysis of soil texture. Soil Sci. Soc.
751 Am. J. 48, 142-147.

752 Shwetha, P., Varija, K., 2015. Soil water retention curve from saturated hydraulic conductivity
753 for sandy loam and loamy sand textured soils. Aquat. Procedia 4, 1142-1149.

754 Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed
755 flexibility in data collection and analysis allows presenting anything as significant.
756 Psychol. Sci. 22, 1359-1366.

757 Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019.
758 Mapping soil hydraulic properties using random forest based pedotransfer functions and
759 geostatistics. Hydrol. Earth Syst. Sci. 23, 2615-2635.

760 Tietje, O., Tapkenhinrichs, M., 1993. Evaluation of pedo-transfer functions. Soil Sci. Soc. Am. J.
761 57, 1088-1095.

762 Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited
763 data in Brazilian Amazonia. Soil Sci. 163, 190-202.

764 Tomasella, J., Hodnett, M.G., Rossato, L., 2000. Pedotransfer functions for the estimation of soil
765 water retention in Brazilian soils. Soil Sci. Soc. Am. J. 64, 327-338.

766 Tóth, B., Makó, A., Toth, G., 2014. Role of soil properties in water retention characteristics of
767 main Hungarian soil types. J. Cent. Eur. Agric. 15, 137-153.

768 Touil, S., Degre, A., Chabaca, M.N., 2016. Sensitivity analysis of point and parametric
769 pedotransfer functions for estimating water retention of soils in Algeria. Soil 2, 647.

770 Twarakavi, N.K., Šimůnek, J., Schaap, M., 2009. Development of pedotransfer functions for
771 estimation of soil hydraulic parameters using support vector machines. Soil Sci. Soc. Am.
772 J. 73, 1443-1452.

773 Ungaro, F., Calzolari, C., Busoni, E., 2005. Development of pedotransfer functions using a group
774 method of data handling for the soil of the Pianura Padano–Veneta region of North Italy:
775 water retention properties. Geoderma 124, 293-317.

776 van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of
777 unsaturated soils. *Soil Sci. Soc. Am. J.* 44, 892-898.

778 Verhagen, J., 1997. Site specific fertiliser application for potato production and effects on N-
779 leaching using dynamic simulation modelling. *Agric. Ecosyst. Environ.* 66, 165-175.

780 Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: A survey
781 and results of new tests. *Pattern Recognit.* 44, 330-349.

782 Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil
783 organic matter, and a proposed modification of the chromic acid titration method. *Soil*
784 *Sci.* 37, 29-38.

785 Wassar, F., Gandolfi, C., Rienzner, M., Chiaradia, E.A., Bernardoni, E., 2016. Predicted and
786 measured soil retention curve parameters in Lombardy region north of Italy. *International*
787 *Soil and Water Conservation Research* 4, 207-214.

788 Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic
789 matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant Soil*
790 340, 7-24.

791 Williams, J., Ross, P., Bristow, K.L., 1992. Prediction of the Campbell water retention function
792 from texture, structure, and organic matter. In 'Indirect methods for estimating the
793 hydraulic properties of unsaturated soils.' University of California: Riverside.

794 Wösten, J., Pachepsky, Y.A., Rawls, W., 2001. Pedotransfer functions: bridging the gap between
795 available basic soil data and missing soil hydraulic characteristics. *J. Hydrol.* 251, 123-
796 150.

797 Zaklouta, F., Stanculescu, B., 2012. Real-time traffic-sign recognition using tree classifiers.
798 *IEEE Transactions on Intelligent Transportation Systems* 13, 1507-1514.

799 Zhao, P., Su, X., Ge, T., Fan, J., 2016. Propensity score and proximity matching using random
800 forest. *Contemp. Clin. Trials* 47, 85-92.

801

802 **Figure captions**

803 ~~Fig. 1. Correlation matrix plot between input and output variables.~~

804 ~~** Correlation is significant at the $P < 0.01$ level.~~

805 ~~* Correlation is significant at the $P < 0.05$ level.~~

806 ~~A list of abbreviations is available in the notation box.~~

807 **Fig 21.** Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten
808 model parameters (θ_r , θ_s , α and n) of the soil water retention curve (SWRC). A list of
809 abbreviations is available in the notation box.

810 **Fig. 32.** An architecture of a random forest.

811 **Fig. 43.** Variation of soil texture classes for the dataset ($n = 223$) on the United States
812 Department of Agriculture (USDA) textural triangle.

813 ~~Fig. 14. Correlation matrix plot between input and output variables.~~

814 ~~** Correlation is significant at the $P < 0.01$ level.~~

815 ~~* Correlation is significant at the $P < 0.05$ level.~~

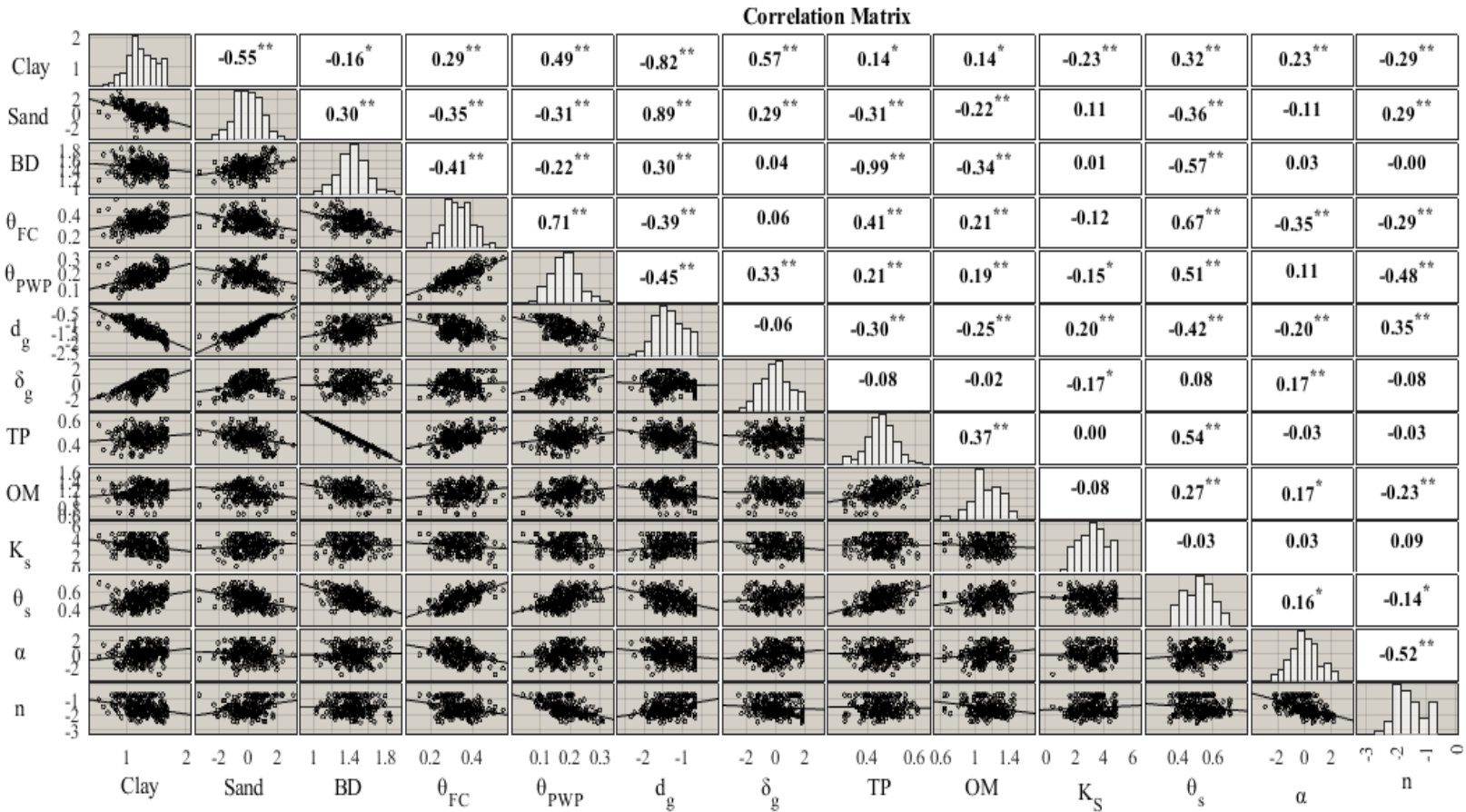
816 ~~A list of abbreviations is available in the notation box.~~

817 **Fig. 5.** Results of the prediction of the soil water retention curve (SWRC) through the van
818 Genuchten model by the non-linear regression (NLR) and random forests (RF) techniques for the
819 training step as reflected in the integral mean error (IME), integral root mean square error
820 (IRMSE), coefficient of determination (R_2), and Akaike's information criterion (AIC). Vertical
821 lines indicate the standard deviations. Means with the same letter are not significantly different at
822 the significance level of $P < 0.05$ (IRMSE only).

823 **Fig. 6.** Results of the prediction of the soil water retention curve (SWRC) through the van
824 Genuchten model by the Rosetta software, non-linear regression (NLR) and random forests (RF)

825 techniques for the testing step as reflected in the integral mean error (*IME*), integral root mean
826 square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion
827 (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not
828 significantly different at the significance level of $P < 0.05$ (*IRMSE* only).
829

830



831

832

Fig. 1. Correlation matrix plot between input and output variables.

833

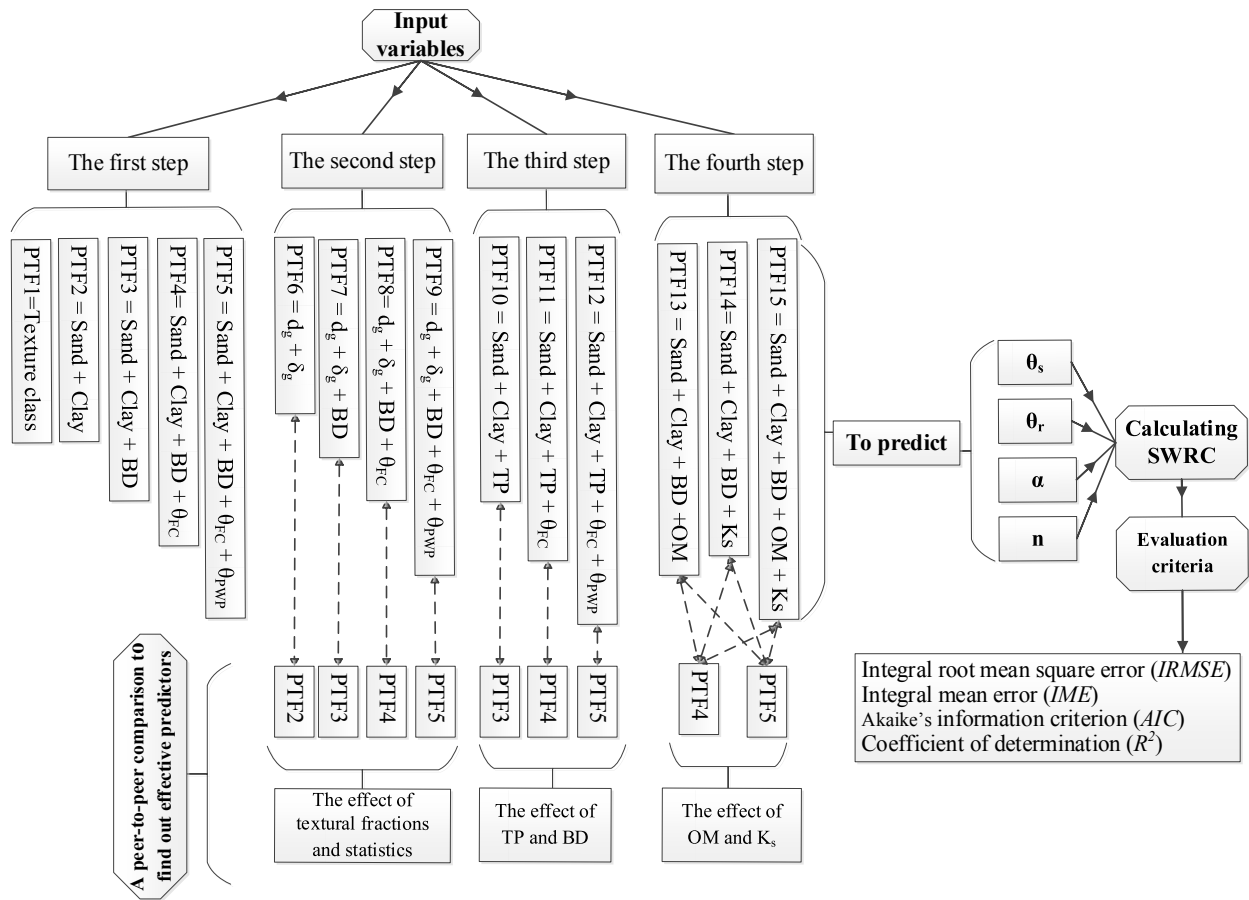
**** Correlation is significant at the 0.01 level.**

834

*** Correlation is significant at the 0.05 level.**

835

A list of abbreviations is available in the notation box.



837
838

839 **Fig 21.** Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten
840 model parameters (θ_r , θ_s , α and n) of the soil water retention curve (SWRC). A list of
841 abbreviations is available in the notation box.

842

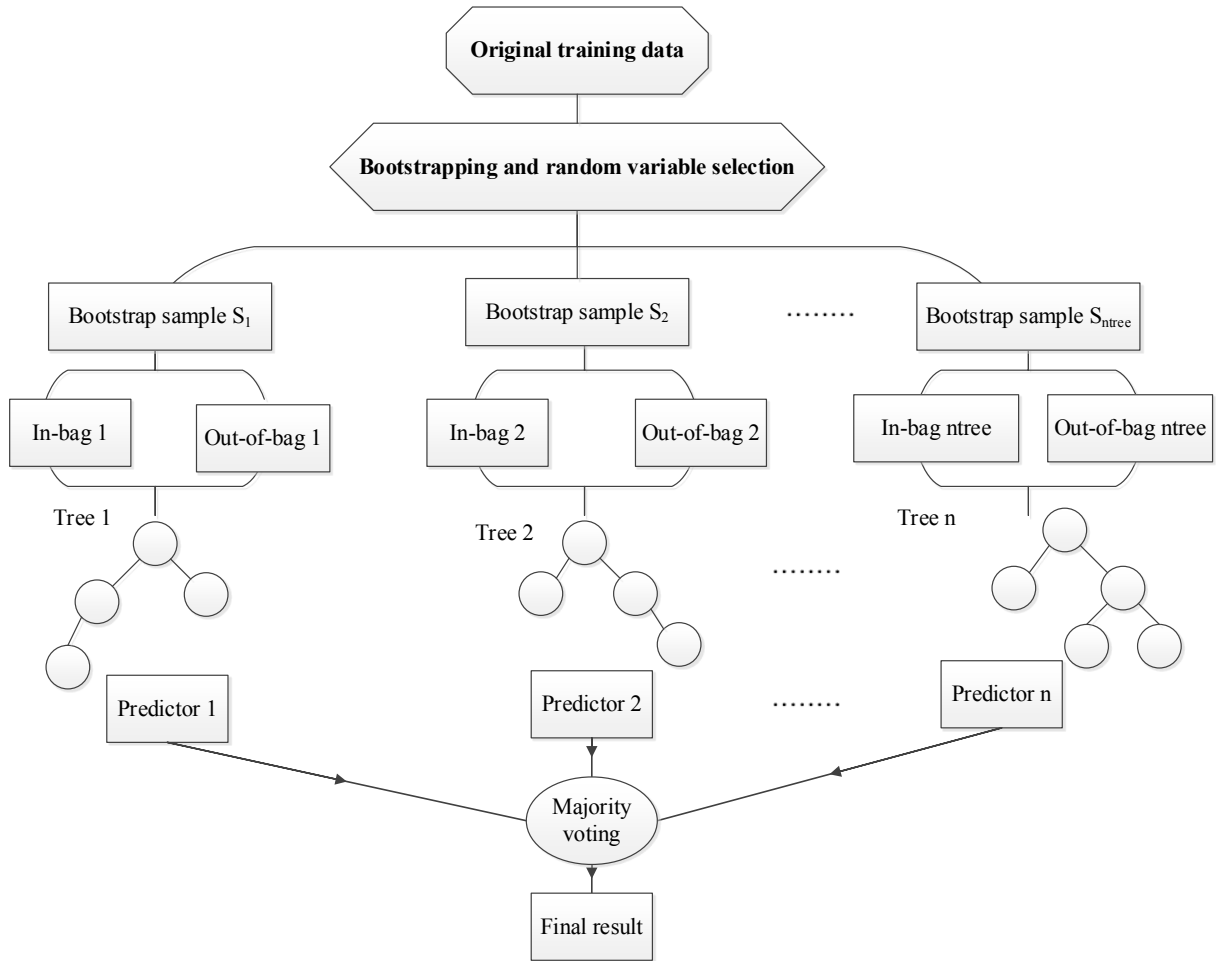
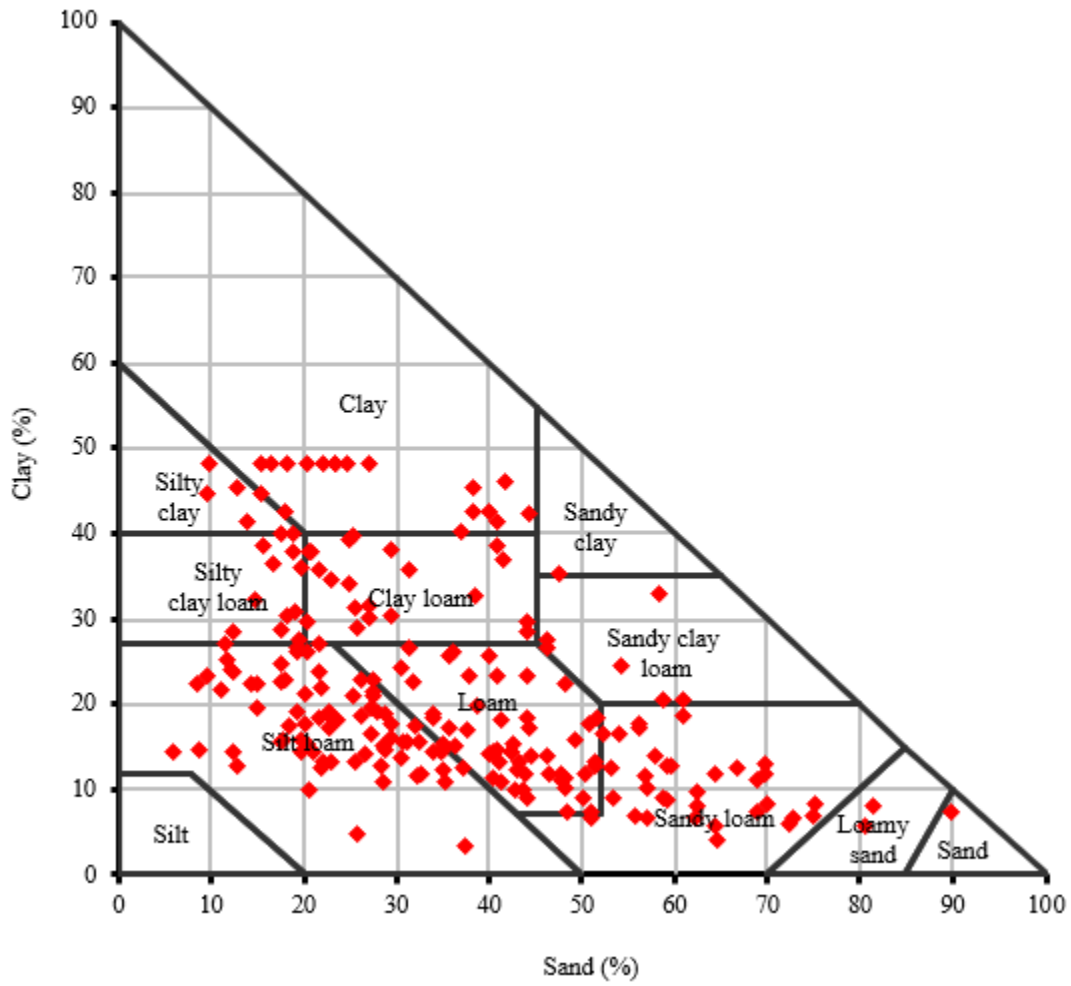


Fig. 32. An architecture of a random forest.



849
850 **Fig. 43.** Variation of soil texture classes for the dataset (n = 223) on the United States
851 Department of Agriculture (USDA) textural triangle.

852

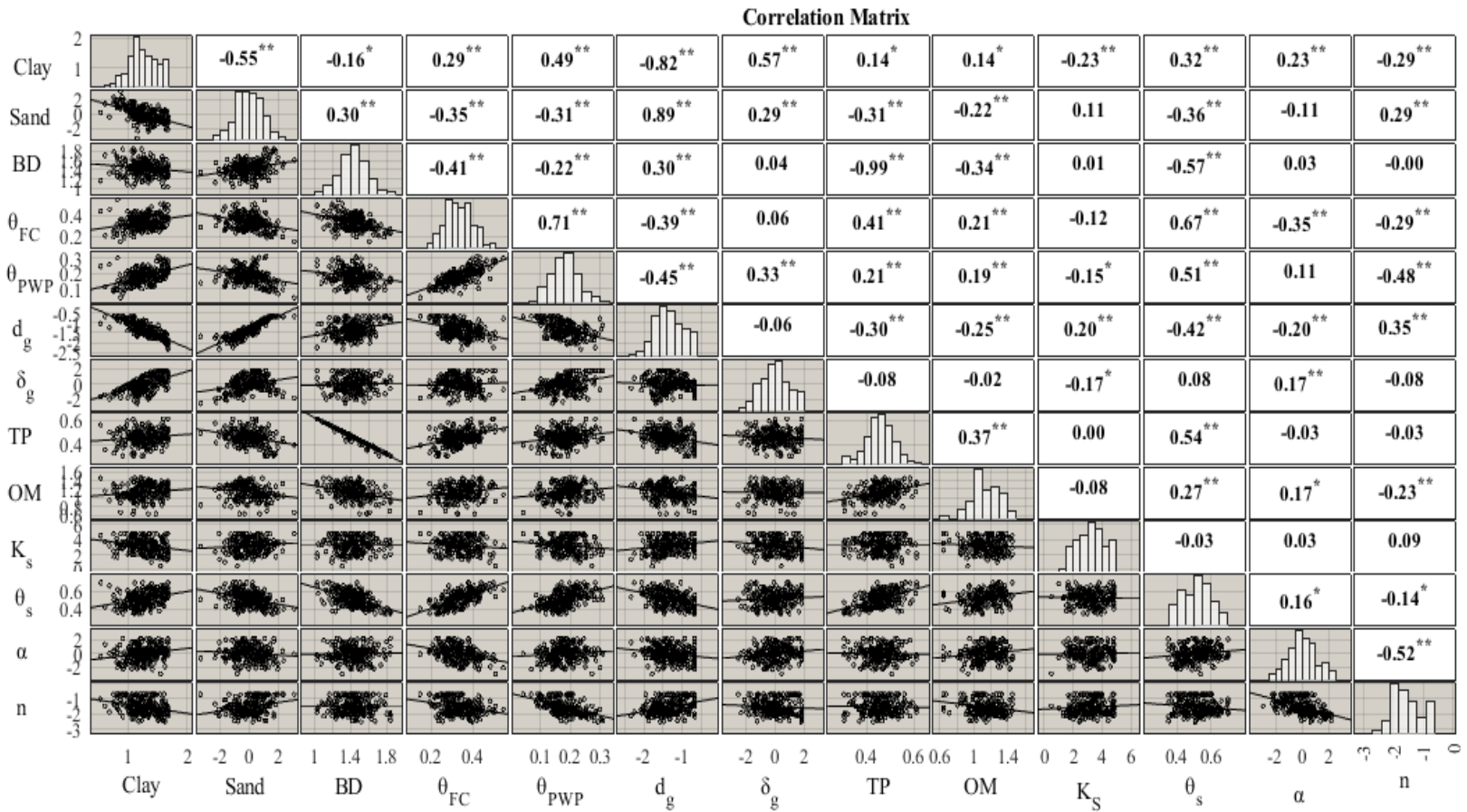
853

854

855

856

857

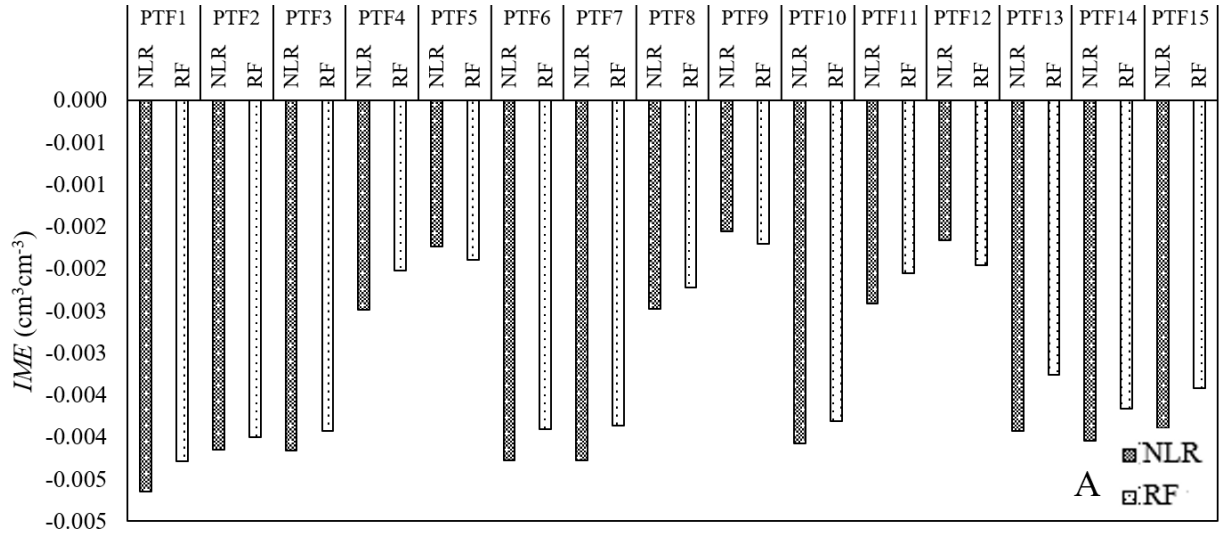


858
859 **Fig. 14.** Correlation matrix plot between input and output variables.

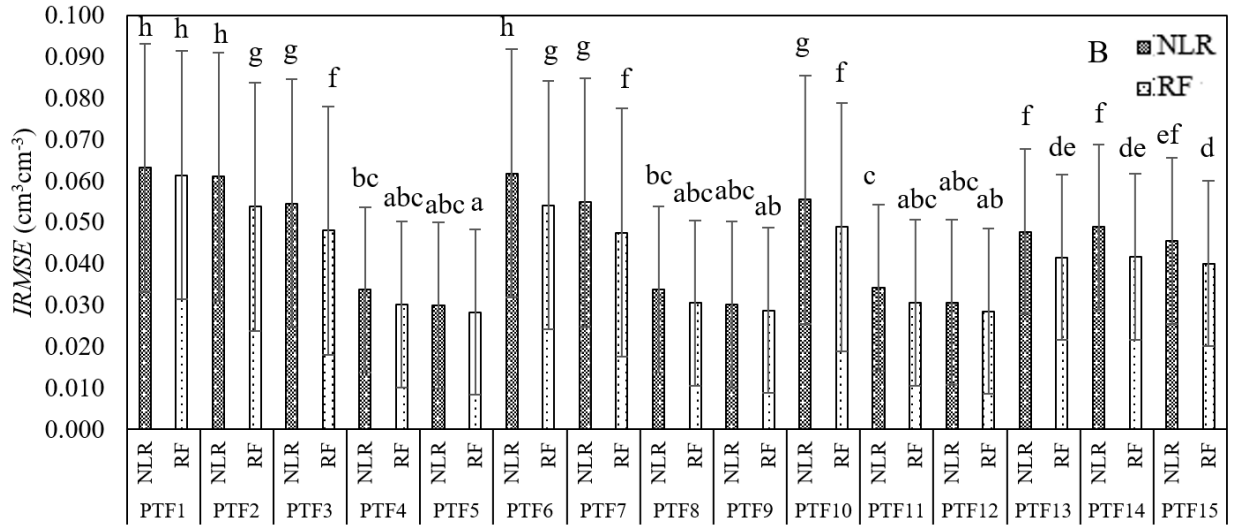
860 ****** Correlation is significant at the $P < 0.01$ level.

861 ***** Correlation is significant at the $P < 0.05$ level.

862 A list of abbreviations is available in the notation box.



863



864

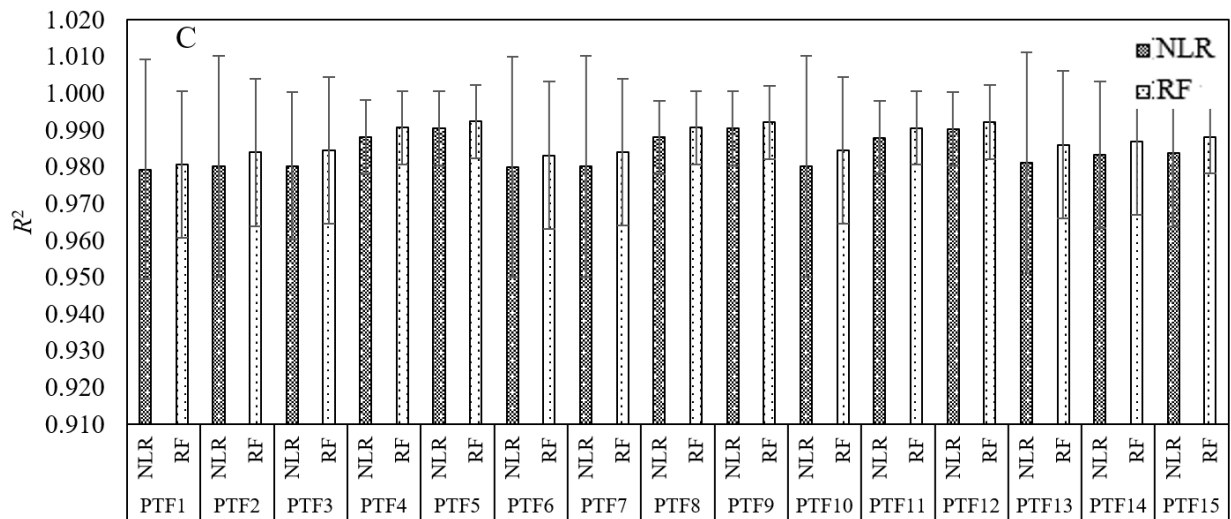
865

866

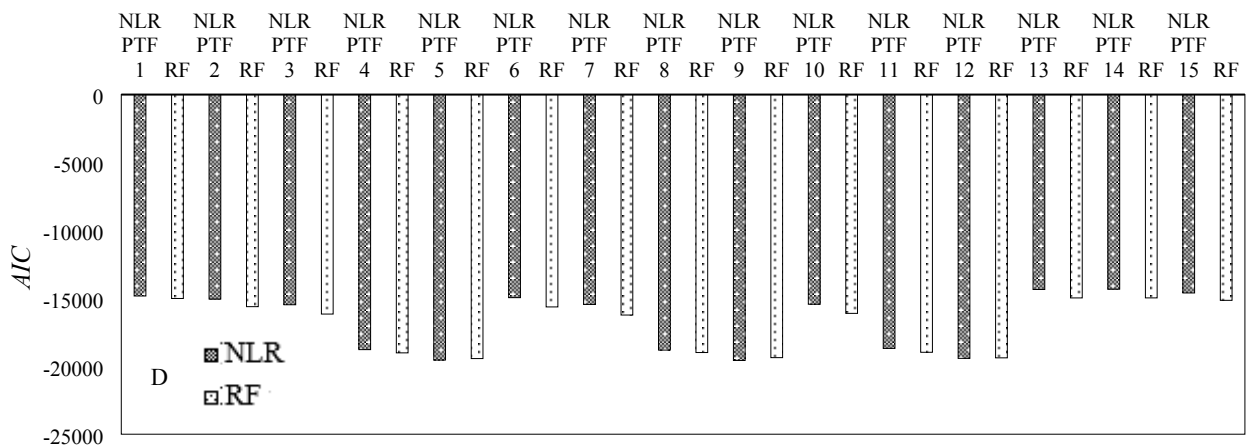
867

868

869



870



871

872

873

874

875

876

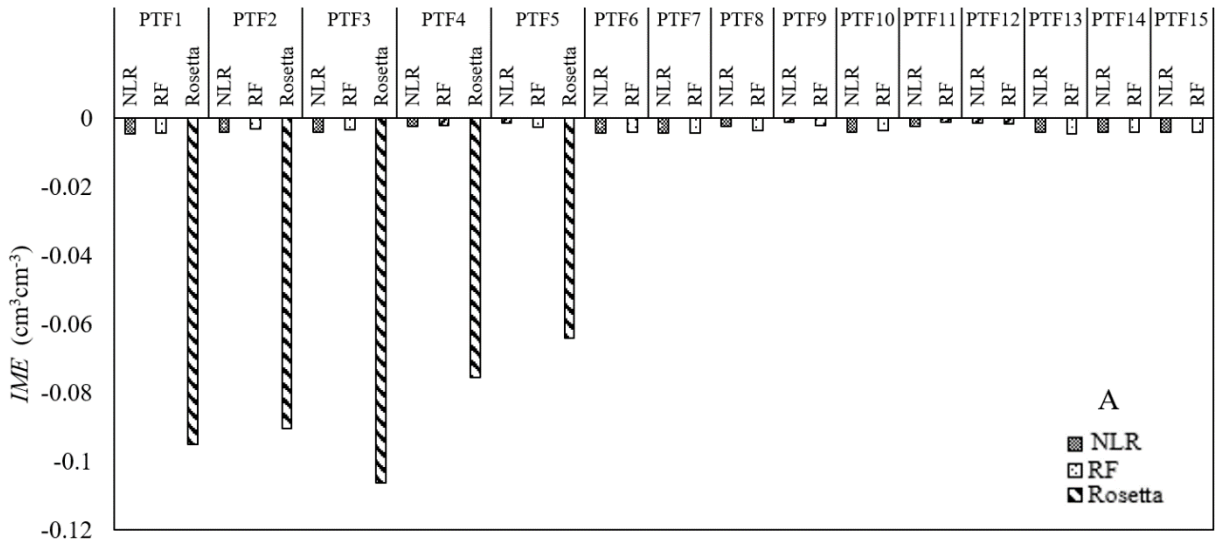
877

878

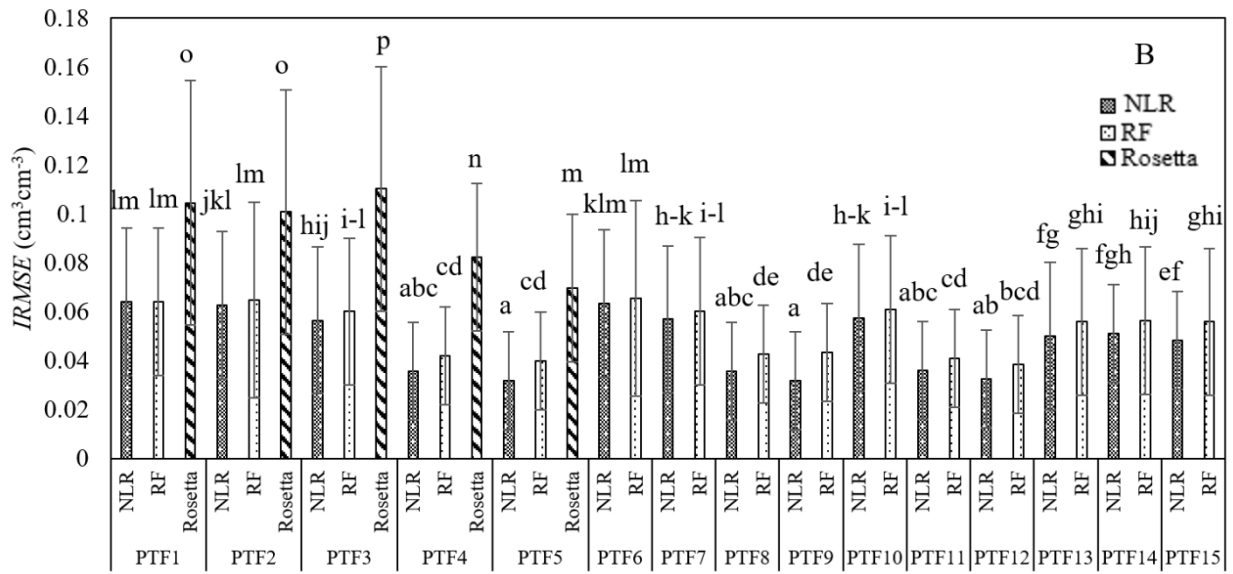
879

Fig. 5. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the non-linear regression (NLR) and random forests (RF) techniques for the training step as reflected in the integral mean error (IME), integral root mean square error (IRMSE), coefficient of determination (R^2), and Akaike's information criterion (AIC). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of $P < 0.05$ (IRMSE only).

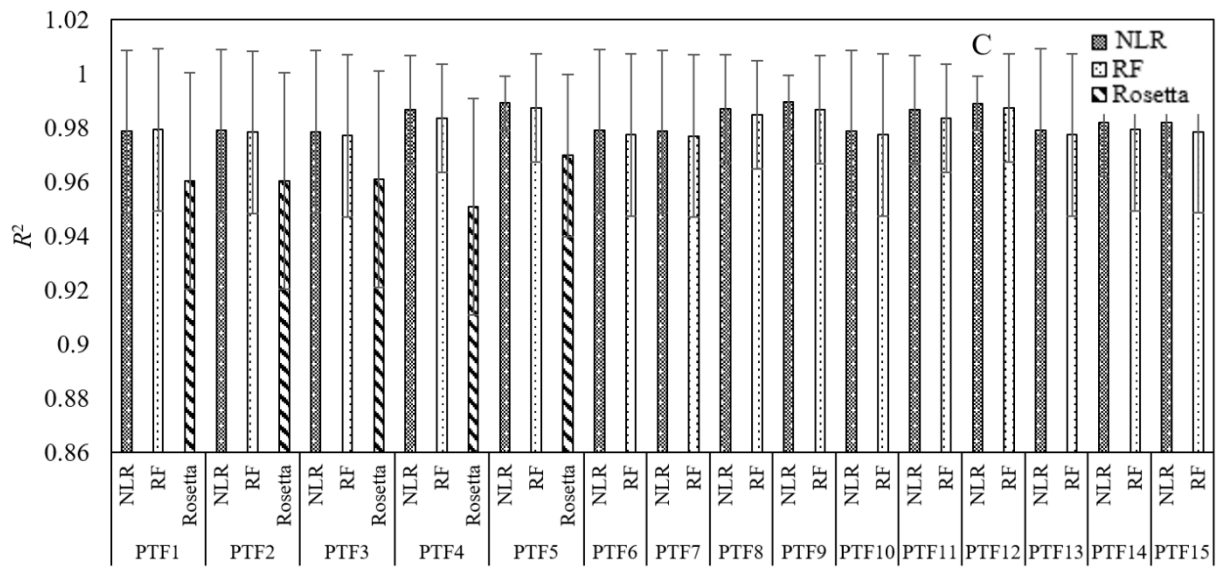
880



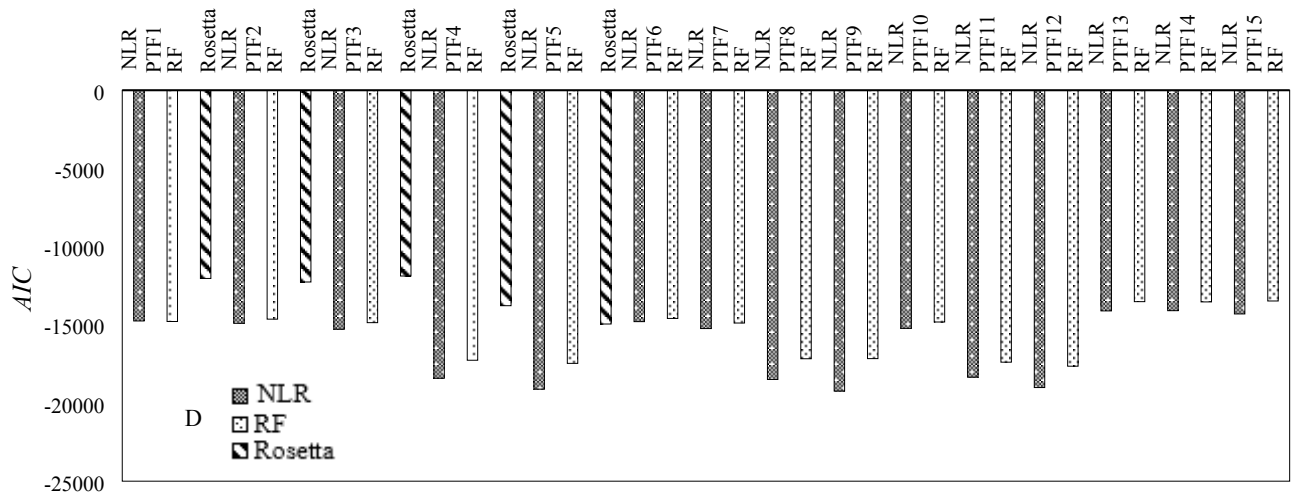
881



882



883



884

885

Fig. 6. Results of the prediction of the soil water retention curve (SWRC) through the van

886

Genuchten model by the Rosetta software, non-linear regression (NLR) and random forests (RF)

887

techniques for the testing step as reflected in the integral mean error (IME), integral root mean

888

square error (IRMSE), coefficient of determination (R_2), and Akaike's information criterion

889

(AIC). Vertical lines indicate the standard deviations. Means with the same letter are not

890

significantly different at the significance level of $P < 0.05$ (IRMSE only).

891

892 **Table 1-** The results of 10, 15 and 20-fold cross-validation (k) for van Genuchten model
 893 parameters of the soil water retention curve derived from nonlinear regression (NLR) and
 894 random forest (RF) techniques based on root mean square error (*RMSE*) for pedotransfer
 895 functions PTF 3, 5 and 11 in the train and test datasets.

			θ_r			θ_s			α			n			
			<i>RMSE</i>			<i>RMSE</i>			<i>RMSE</i>			<i>RMSE</i>			
			Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	
PTF3	k=10	NLR	0.058	0.060	0.059	0.063	0.065	0.064	1.017	1.037	1.027	0.426	0.436	0.431	
		RF	0.052	0.061	0.056	0.058	0.073	0.066	0.893	1.084	0.989	0.374	0.442	0.408	
	k=15	NLR	0.058	0.060	0.059	0.064	0.064	0.064	1.017	1.030	1.024	0.426	0.434	0.430	
		RF	0.052	0.061	0.057	0.058	0.070	0.064	0.894	1.033	0.964	0.374	0.441	0.408	
	k=20	NLR	0.058	0.060	0.059	0.064	0.064	0.064	0.064	0.064	0.064	0.064	0.426	0.437	0.432
		RF	0.051	0.060	0.056	0.057	0.071	0.064	0.057	0.071	0.064	0.368	0.442	0.405	
PTF5	k=10	NLR	0.051	0.053	0.052	0.053	0.054	0.054	0.764	0.796	0.780	0.380	0.397	0.389	
		RF	0.043	0.056	0.050	0.046	0.056	0.051	0.675	0.869	0.772	0.327	0.411	0.369	
	k=15	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.764	0.790	0.777	0.381	0.399	0.390	
		RF	0.044	0.054	0.049	0.046	0.055	0.050	0.679	0.848	0.763	0.329	0.421	0.375	
	k=20	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.765	0.789	0.777	0.381	0.399	0.390	
		RF	0.042	0.054	0.048	0.044	0.054	0.049	0.654	0.842	0.748	0.316	0.412	0.364	
PTF11	k=10	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.018	1.052	1.035	0.431	0.448	0.440	
		RF	0.050	0.061	0.056	0.047	0.057	0.052	0.770	0.978	0.874	0.370	0.443	0.406	
	k=15	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.019	1.037	1.028	0.432	0.447	0.439	
		RF	0.050	0.060	0.055	0.047	0.057	0.052	0.770	1.009	0.889	0.369	0.450	0.410	
	k=20	NLR	0.058	0.060	0.059	0.065	0.065	0.065	1.020	1.024	1.022	0.432	0.439	0.435	
		RF	0.049	0.061	0.055	0.046	0.056	0.051	0.745	0.964	0.855	0.361	0.443	0.402	

896

897 **Table 12-** Some descriptive statistics of the measured soil variables and parameters of the van
 898 Genuchten model of the soil water retention curve for the entire dataset (223 soil samples).

Variables ^a	Mean	CV (%)	Minimum	Maximum	<i>P</i> -value
Clay content (%)	21.39	54.05	3.47	48.00	0.00
Log (clay content)	1.27	19.08	0.54	1.68	0.66
Sand content (%)	35.45	48.40	5.90	89.80	0.00
Sand content*	-0.01	-14350.94	-3.40	3.14	0.90
Bulk density (g cm ⁻³)	1.43	10.97	1.03	1.84	0.83
θ_{FC} (cm ³ cm ⁻³) [§]	0.33	20.44	0.15	0.55	0.45
θ_{PWP} (cm ³ cm ⁻³)	0.18	26.21	0.04	0.31	0.90
d_g (mm)	0.07	86.62	0.00	0.21	0.00
Log (d_g)	-1.33	-27.91	-2.34	-0.67	0.77
δ_g (-)	11.57	29.39	4.54	19.97	0.00
δ_g^*	-0.01	-9872.87	-2.53	1.80	0.96
Total porosity (cm ³ cm ⁻³)	0.46	13.26	0.31	0.61	0.67
Organic matter content (%)	1.84	53.68	0.17	4.41	0.00
(Organic matter content) ^(1/4)	1.13	14.83	0.64	1.45	0.86
K_s (cm day ⁻¹)	169.10	96.58	0.06	530	0.00
$(K_s)^{(1/4)}$	3.23	30.37	0.50	4.80	0.59
θ_r (cm ³ cm ⁻³)	0.04	158.05	0.00	0.17	0.00
θ_s (cm ³ cm ⁻³)	0.52	16.26	0.35	0.70	0.56
α (kPa ⁻¹)	0.06	115.62	0.00	0.29	0.00
α^*	0.01	8889.14	-2.93	2.19	0.93
n	1.24	9.80	1.08	1.48	0.00
Ln ($n-1$)	-1.55	-30.92	-2.52	-0.74	0.05

899 ^a CV, coefficient of variation.

900 [§]. A list of abbreviations is available in the notation box.

901 * Normalized form of sand content: $0.91+1.06 \times \text{Ln}((\text{sand content}- 4.3)/(100.2-\text{sand content}))$;

902 normalized form of δ_g : $-1.04657+1.39359 \times \text{Asinh}((\delta_g- 8.4)/3.04)$; and normalized form of α :

903 $3.6+0.92 \times \text{Ln}((\alpha- 8.2 \times 10^{-6})/(1.6-\alpha))$. *P*-value is a significance value for normality test.

904

905 **Table 23-** The variance inflation factor (*VIF*) values for normalized form of the input variables.

<u>PTFs</u>	<u>Clay* (%)</u>	<u>Sand (%)</u>	<u>BD^s (g cm⁻³)</u>	<u>θ_{FC} (cm³ cm⁻³)</u>	<u>θ_{PWP} (cm³ cm⁻³)</u>	<u>d_p (mm)</u>	<u>δ_g (-)</u>	<u>TP (cm³ cm⁻³)</u>	<u>OM (%)</u>	<u>K_s (cm day⁻¹)</u>
<u>PTF2</u>	<u>1.42</u>	<u>1.42</u>								
<u>PTF3</u>	<u>1.43</u>	<u>1.52</u>	<u>1.10</u>							
<u>PTF4</u>	<u>1.45</u>	<u>1.56</u>	<u>1.25</u>	<u>1.31</u>						
<u>PTF5</u>	<u>1.79</u>	<u>1.58</u>	<u>1.27</u>	<u>2.48</u>	<u>2.56</u>					
<u>PTF6</u>						<u>1.00</u>	<u>1.00</u>			
<u>PTF7</u>			<u>1.11</u>			<u>1.11</u>	<u>1.01</u>			
<u>PTF8</u>			<u>1.25</u>	<u>1.33</u>		<u>1.01</u>	<u>1.22</u>			
<u>PTF9</u>			<u>1.28</u>	<u>2.50</u>	<u>2.73</u>	<u>1.34</u>	<u>1.22</u>			
<u>PTF10</u>	<u>1.55</u>	<u>1.43</u>						<u>1.11</u>		
<u>PTF11</u>	<u>1.58</u>	<u>1.46</u>		<u>1.32</u>				<u>1.26</u>		
<u>PTF12</u>	<u>1.60</u>	<u>1.79</u>		<u>2.49</u>	<u>2.56</u>			<u>1.28</u>		
<u>PTF13</u>	<u>1.48</u>	<u>1.65</u>	<u>1.25</u>						<u>1.14</u>	
<u>PTF14</u>	<u>1.55</u>	<u>1.64</u>	<u>1.14</u>							<u>1.06</u>
<u>PTF15</u>	<u>1.55</u>	<u>1.65</u>	<u>1.25</u>						<u>1.15</u>	<u>1.06</u>

906 * Normalized form of the input variables is available in Table 2.

907 ^s. A list of abbreviations is available in the notation box.

908

909 **Table 34-** Analysis of variance of the integral root mean square error (IRMSE) of the prediction
 910 of the soil water retention curveSWRC by different methods (nonlinear regression and random
 911 forest) and pedotransfer functions (PTFs 1-15) for both the train and test datasets.

	<u>Source</u>	<u>Degree freedom</u>	<u>Mean square</u>	<u>F-value</u>	<u>P-value</u>
<u>Train</u>	<u>Repeat (Block)</u>	<u>222</u>	<u>0.007</u>	<u>19.09</u>	<u><0.0001</u>
	<u>PTFs</u>	<u>14</u>	<u>0.062</u>	<u>180.68</u>	<u><0.0001</u>
	<u>Methods</u>	<u>1</u>	<u>0.038</u>	<u>109.69</u>	<u><0.0001</u>
	<u>PTFs × Methods</u>	<u>14</u>	<u>0.001</u>	<u>1.78</u>	<u>0.0356</u>
	<u>Error</u>	<u>6288</u>	<u>0.0003</u>		
<u>Test</u>	<u>Repeat (Block)</u>	<u>222</u>	<u>0.010</u>	<u>16.04</u>	<u><0.0001</u>
	<u>PTFs</u>	<u>14</u>	<u>0.073</u>	<u>117.22</u>	<u><0.0001</u>
	<u>Methods</u>	<u>2</u>	<u>0.656</u>	<u>1056.43</u>	<u><0.0001</u>
	<u>PTFs × Methods</u>	<u>18</u>	<u>0.002</u>	<u>3.68</u>	<u><0.0001</u>
	<u>Error</u>	<u>7398</u>	<u>0.0006</u>		

912

913

- The RF was compared to NLR method and Rosetta-based PTFs to predict the SWRC
- The NLR method had better performance due to higher reliability in the testing step
- The RF and NLR-based PTFs performed better than the Rosetta-based PTFs
- In the absence of moisture points, OM and K_s can be suitable predictors for SWRC
- d_g and δ_g can be suitable alternatives for textural fractions in predicting SWRC
- Total porosity and bulk density have the same effect in predicting the SWRC

1 **Estimating the soil water retention curve: comparison of multiple nonlinear regression**
2 **approach and random forest data mining technique**

3 **M. Rastgou¹, H. Bayat^{2*}, and M. Mansoorizadeh³, Andrew S. Gregory⁴**

4
5 ¹ **Mostafa Rastgou:** Ph. D. Student of Soil Science, Department of Soil Science, Faculty of
6 Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: mostafa.rastgo@gmail.com,

7 ² **Hossein Bayat:** Associate Professor (Ph. D.), Department of Soil Science, Faculty of Agriculture,
8 Bu Ali Sina University, Hamedan, Iran. Postal Address: Department of Soil Science, Faculty of
9 Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: h.bayat@basu.ac.ir Other e-mail:
10 hbayat2001@gmail.com. Office phone: +98-81-34424189, Mobile phone: +98-918-8188378.
11 Fax: +98-81-34424189.

12 ³ **Muharram Mansoorizadeh:** Assistant professor (Ph. D.), Department of Computer Science,
13 Faculty of Engineering, Bu Ali Sina University, Hamedan, Iran. E-mail: mansoorm@basu.ac.ir

14 ⁴ **Andrew S. Gregory:** Sustainable Agriculture Sciences Department, Rothamsted Research,
15 Harpenden, Hertfordshire, AL5 2JQ, UK. E-mail: andy.gregory@rothamsted.ac.uk

16
17 *Corresponding author (h.bayat@basu.ac.ir, Other e-mail: hbayat2001@gmail.com).

24 **Estimating the soil water retention curve: comparison of multiple nonlinear regression**
25 **approach and random forest data mining technique**

26 **Abstract**

27 This study evaluates the performance of the random forest (RF) method on the prediction of the
28 soil water retention curve (SWRC) and compares its performance with those of nonlinear
29 regression (NLR) and Rosetta-based pedotransfer functions (PTFs), which has not been reported
30 so far. Fifteen RF and NLR-based PTFs were constructed using readily-available soil properties
31 for 223 soil samples from Iran. The general performance of RF and NLR-based PTFs was
32 quantified by the integral root mean square error (*IRMSE*), Akaike's information criterion (*AIC*)
33 and coefficient of determination (R^2). The results showed that the accuracy of the RF-based PTFs
34 was significantly ($P < 0.05$) better than the NLR-based PTFs, and that the reliability of the NLR-
35 based PTFs was significantly ($P < 0.01$) better than the RF-based PTFs and all of the Rosetta-
36 based PTFs. The average values of the *IRMSE*, *AIC* and R^2 of the RF method were 0.041 cm^3
37 cm^{-3} , -16997.7 , and 0.987 , and $0.053 \text{ cm}^3 \text{ cm}^{-3}$, -15547.5 , and 0.981 for the training and testing
38 steps of all PTFs, respectively, whereas the values for the NLR method were $0.046 \text{ cm}^3 \text{ cm}^{-3}$, $-$
39 16616.4 , and 0.984 , and $0.048 \text{ cm}^3 \text{ cm}^{-3}$, -16355.6 , and 0.983 for the training and testing steps,
40 respectively. The PTF5 of the RF and NLR methods, with inputs of sand and clay contents, bulk
41 density, and the water content at field capacity and permanent wilting point, had the greatest R^2
42 values (0.987 and 0.989 , respectively), and the lowest *IRMSE* values (0.039 and $0.032 \text{ cm}^3 \text{ cm}^{-3}$,
43 respectively) compared to other PTFs for the testing step. Overall, the RF method had less
44 reliability for the prediction of the SWRC compared to the NLR method due to overprediction,
45 uncertainty of determination of forest scale and instability in the testing step. These findings
46 could provide the scientific basis for further research on the RF method.

47 *Keywords:* pedotransfer functions; soil water retention curve; soil texture; soil structure; van
48 Genuchten.
49

Notation	
Sand content (%)	S
Clay content (%)	C
Geometric mean diameter (mm)	d_g
Geometric standard deviation (-)	δ_g
Bulk density (g cm^{-3})	BD
Total porosity ($\text{cm}^3 \text{cm}^{-3}$)	TP
Water content at field capacity, 33 kPa ($\text{cm}^3 \text{cm}^{-3}$)	θ_{FC}
Water content at 1500 kPa ($\text{cm}^3 \text{cm}^{-3}$)	θ_{PWP}
Organic matter content (%)	OM
Saturated hydraulic conductivity (cm day^{-1})	K_s
Saturated water content ($\text{cm}^3 \text{cm}^{-3}$)	θ_s
Residual water content ($\text{cm}^3 \text{cm}^{-3}$)	θ_r
Random forest	RF
Nonlinear regression	NLR
Soil water retention curve	SWRC

50

51 **1 Introduction**

52 Soil hydraulic properties are principle factors that control the movement of water and solutes in
53 the soil. Determination of the soil hydraulic properties is required for many distinct applications
54 linked with irrigation, land use planning, drainage and drought risk assessment (Dobarco et al.,
55 2019). The soil water retention curve (SWRC) is one of the most important soil hydraulic
56 properties. It defines the relationship between soil matric potential and soil water content (Hillel,
57 1998). The SWRC is a crucial parameter in soil and water management for sustainable and

58 improved agricultural production (Shwetha and Varija, 2015). The SWRC depends principally
59 on texture, structure and bulk density (BD) of soils (Wassar et al., 2016). Many methods have
60 been introduced for the direct measurement of the SWRC in the laboratory (e.g., the hanging
61 water column and pressure plate methods) (Klute, 1986) and in the field (e.g., tensiometric)
62 (Bruce and Luxmoore, 1986). Measurements of the SWRC at several matric potentials can be
63 expensive, difficult and time-consuming, hence it is common to predict it by modelling (Dobarco
64 et al., 2019). Modelling of soil water is an essential tool in evaluating the effects of different
65 managements on crop yield and environmental quality (Verhagen, 1997).

66 Pedotransfer functions (PTFs) translate easy-to-measure data that we have (e.g., texture class,
67 particle size distribution (PSD) and BD) into difficult-to-measure data that we need (soil
68 hydraulic data) (Bouma, 1989). Estimates of the SWRC by PTFs are valuable in many studies,
69 such as hydrology, soil mapping and hydrogeology (Børgesen and Schaap, 2005). The point- and
70 parametric-based PTFs are generally developed to predict water content at specific matric
71 potential values and the entire SWRC, respectively, by multiple linear (MLR) and nonlinear
72 regression (NLR) methods (Gunarathna et al., 2019b; Merdun et al., 2006; Minasny et al., 1999;
73 Rajkai et al., 2004; Tomasella et al., 2000). Data mining techniques including artificial neural
74 networks (ANNs) (Bayat et al., 2013a; Bayat et al., 2013b; Gunarathna et al., 2019a; Koekkoek
75 and Booltink, 1999; Pachepsky et al., 1996), group method of data handling (GMDH) (Bayat et
76 al., 2011; Neyshaburi et al., 2015; Pachepsky and Rawls, 1999), nonparametric nearest neighbor
77 technique (Botula et al., 2013; Gunarathna et al., 2019a; Haghverdi et al., 2015; Nemes et al.,
78 2006; Nguyen et al., 2017) and support vector machine (SVM) (Khlosi et al., 2016; Lamorski et
79 al., 2008; Lamorski et al., 2014; Twarakavi et al., 2009), have been applied successfully for PTF
80 development.

81 Random forest (RF), or random decision forests, has become a popular approach as an ensemble
82 learning method for prediction and classification (Verikas et al., 2011). The RF method has been
83 developed by Breiman (2001) as an expansion of the classification and regression trees (CART)
84 technique to provide better performance of prediction results (Wiesmeier et al., 2011). So far,
85 few studies have been carried out on the application of the RF method in soil science. Tóth et al.
86 (2014) applied the RF method to analyze the relationship between soil water content at four
87 matric suctions (0.1, 33, and 1500 kPa, and 150 MPa) and Hungarian soil map information. They
88 found that the importance of soil properties in the prediction of the soil water content varied
89 according to soil type and matric suction. Recently Szabó et al. (2019) have developed PTFs
90 based on RF and geostatistics methods to map soil hydraulic properties, such as water contents at
91 saturation, field capacity and wilting point, for the Balaton catchment area in Hungary. Araya
92 and Ghezzehei (2019) compared the performances of four machine-learning algorithms including
93 the k-nearest neighbors (kNNs), support vector regression (SVR), RF, and boosted regression
94 tree (BRT) for prediction of saturated hydraulic conductivity. They found that the BRT model
95 outperformed the other algorithms closely followed by the RF model. Gunarathna et al. (2019a)
96 tested three machine-learning algorithms including ANN, kNN, and RF to estimate volumetric
97 water content at matric suctions of 10, 33 and 1500 kPa for soils in Sri Lanka. They
98 recommended that the PTFs to be developed using the RF algorithm. Ließ et al. (2012) studied
99 uncertainty in the spatial prediction of soil texture by comparison of the RF and regression tree
100 techniques for 56 soil profiles and found that the former method provided a better result. Also,
101 Wiesmeier et al. (2011) utilized the RF technique to develop digital mapping of the soil organic
102 matter content in 120 soil profiles. They found that the prediction accuracy of the RF modeling
103 was acceptable. A review of literatures therefore revealed that the RF data mining technique has

104 been applied to develop PTFs to predict specific points of the SWRC, such as field capacity and
105 permanent wilting point, or particular properties such as saturated hydraulic conductivity, but it
106 has not been used to develop parametric-based PTFs of the van Genuchten model parameters, so
107 far. Therefore, the objective of the present study was to develop simple parametric-PTFs to
108 predict the SWRC with greater accuracy and reliability using a novel approach with the RF data
109 mining technique. We compare its performance with those of the multiple NLR approach and
110 with Rosetta software (Schaap et al., 2001) on the prediction of the SWRC through finding the
111 best input variables and PTFs for the SWRC.

112

113 **2 Materials and methods**

114 *2.1 Sample collection and determination*

115 In the present study 223 undisturbed and disturbed soil samples were taken from six provinces of
116 Iran including west Azarbaijan ($35^{\circ} 8' - 39^{\circ} 46' N$, $44^{\circ} 3' - 47^{\circ} 23' E$; 60 data), Hamedan
117 ($33^{\circ} 59' - 35^{\circ} 48' N$, $47^{\circ} 34' - 49^{\circ} 36' E$; 55 data), Kermanshah ($33^{\circ} 41' - 35^{\circ} 17' N$, 45°
118 $24' - 48^{\circ} 6' E$; 26 data), Kurdistan ($34^{\circ} 45' - 36^{\circ} 31' N$, $45^{\circ} 31' - 48^{\circ} 13' E$; 22 data),
119 Mazandaran ($35^{\circ} 46' - 36^{\circ} 58' N$, $50^{\circ} 21' - 58^{\circ} 08' E$; 30 data) and Fars ($27^{\circ} 2' - 31^{\circ} 42'$
120 N , $50^{\circ} 42' - 55^{\circ} 38' E$; 30 data). Steel cylinders, measuring 5.1 cm in diameter and 3.5 cm in
121 height, were used to collect the undisturbed samples. Since the sampling was done from different
122 locations of the various provinces, the topsoil and subsoil layers of soil at different locations had
123 different depths and thicknesses. We collected samples from the center of the topsoil and subsoil
124 layers, which represented the pedological A and B horizons, respectively. The sampling depths
125 varied from 10 to 35 cm for topsoil (208 samples) and from 20 to 45 cm for subsoil (15 samples),
126 reflecting the variation in the soil profiles.

127 Soil PSD was analyzed by the hydrometer method (Gee and Or, 2002), and the geometric mean
 128 and standard deviation of particle diameter (d_g and δ_g , respectively) were calculated by
 129 equations from Shirazi and Boersma (1984). Organic matter (OM) content was determined by
 130 the Walkley and Black (1934) method and BD by the core method (Blake and Hartge, 1986).
 131 Total porosity (TP) was calculated from BD and particle density, and the saturated hydraulic
 132 conductivity (K_s) was measured with a constant head permeameter (Klute and Dirksen, 1986).
 133 The SWRC was constructed by measuring the volumetric water content at matric suctions of 0
 134 (saturation status of soil samples), 1, 2 and 5 kPa with a sandbox apparatus, and at 10, 25, 50,
 135 100, 200, 500, 1000 and 1500 kPa with a pressure plate apparatus. Undisturbed samples were
 136 used for measurement of the matric suctions from 0 to 100 kPa and disturbed samples were used
 137 for matric suctions from 200 to 1500 kPa. Two key points in the SWRC are the water contents at
 138 field capacity (30 kPa suction; θ_{FC}) and permanent wilting point (1500 kPa suction; θ_{PWP}).

139

140 2.2 Soil-water retention equation

141 The van Genuchten–Mualem (Eq. (1)) model (Mualem, 1976; van Genuchten, 1980) was utilized
 142 to describe the SWRC data.

$$\theta = \theta_r + (\theta_s - \theta_r) \times \frac{1}{\left[1 + (\alpha h)^n\right]^{\left(\frac{1-1/n}{n}\right)}} \quad (1)$$

143 where θ_r and θ_s are residual and saturated water contents ($\text{cm}^3 \text{ cm}^{-3}$), respectively, and h is the
 144 soil water suction (kPa). The parameter α is related to the inverse of the air entry pressure (>0 ,
 145 kPa^{-1}) and n (>1 , dimensionless parameter) is related to the pore size distribution of the soil (van
 146 Genuchten, 1980). In the present study, van Genuchten model parameters θ_r , θ_s , α and n were
 147 obtained using the MATLAB software (MathWorks, 2018).

149 2.3 Data pre-processing

150 Data pre-processing and regression assumptions, including detection of outliers, normality test of
151 the residuals, multicollinearity and independence of the residuals, were applied for all variables
152 (Berry, 1993). The outliers in the data were identified by the inter-quartile range (IQR) method
153 (Seo, 2006) and were replaced by the lower and upper threshold values (MathWorks, 2018).
154 Before developing PTFs, all variables were evaluated by Kolmogorov-Smirnov normality and
155 multicollinearity tests by the SPSS 24 software (IBM, 2016). The degree of multicollinearity in
156 the PTFs was tested by the variance inflation factor ($VIF=1/1-R^2_j$, where R^2_j is the R^2 value
157 obtained by regressing the j^{th} predictor on the remaining predictors) (Hocking, 2013). Also, to
158 avoid multicollinearity between textural contents, the silt fraction was not used as a predictor.
159 The variables clay content, sand content, d_g , δ_g , OM, K_s , α and n had non-normal distributions,
160 therefore, transformations were applied to normalize them.

161

162 2.4 Developing PTFs

163 The PTF inputs were arranged in four steps (Fig. 1). The first step (PTFs 1-5) was based on basic
164 soil properties (i.e., sand content (%), clay content (%), BD (g cm^{-3}), θ_{FC} ($\text{cm}^3 \text{cm}^{-3}$) and θ_{PWP}
165 ($\text{cm}^3 \text{cm}^{-3}$)) according to Rosetta-based PTFs (Schaap et al., 2001) for comparison of SWRC
166 estimates by other methods. The parameters of the van Genuchten model were predicted in all
167 steps. In the second step (PTFs 6-9), d_g (mm) and δ_g were used as new inputs instead of sand and
168 clay contents in the previous step to evaluate the efficiency of using statistical descriptors of PSD
169 to predict the parameters of the van Genuchten model. To build the third step (PTFs 10-12), TP
170 ($\text{cm}^3 \text{cm}^{-3}$) replaced BD from PTFs 3-5 to evaluate the effect of using TP instead of BD on the
171 prediction of the parameters of the van Genuchten model. In other words, the purpose of the

172 second and third steps was to evaluate whether the use of another form of descriptors of soil
173 structure (TP instead of the BD) and soil texture (d_g and δ_g instead of the sand and clay contents)
174 would improve the accuracy of the estimates or not. In the last step, PTFs 13-15 were developed
175 by including OM (%) and K_s (cm day^{-1}) as new variables to evaluate the efficiency of these
176 instead of the water content at specific matric suctions on the prediction of the van Genuchten
177 model parameters. The input variables of the 15 PTFs are shown in Fig. 1.

178 To compare the results of PTFs 1-5 of the RF and NLR methods with those of the Rosetta
179 models, the parameters of the van Genuchten model (θ_r , θ_s , α and n) were estimated by the PTFs
180 built in the Rosetta software (PTFs 1-5), using the measured values of input variables based on
181 PTFs 1-5 as predictors in the Rosetta program. The estimated coefficients of the van Genuchten
182 model were used to calculate the estimated water content at matric suctions from 0 to 1500 kPa
183 (estimated SWRCs). Then curve-by-curve comparison of the measured and estimated SWRCs
184 was performed with different evaluation statistics. Since there is no training step in the Rosetta
185 software, the results of the Rosetta model was only compared with the results of the testing step.

186 To evaluate the effect of using different descriptors of PSD on the prediction of the SWRC, PTFs
187 6, 7, 8 and 9 from the second step were compared with PTFs 2, 3, 4 and 5 from the first step,
188 respectively (Fig. 1). In the same way, to evaluate effect of using different descriptors of soil
189 structure on the prediction of the SWRC, PTFs 10, 11 and 12 from the third step were compared
190 with PTFs 3, 4 and 5 from the first step, respectively. Also, the PTFs 13-15 were compared with
191 the PTFs 4 and 5 to find out the efficiency of OM and K_s variables as predictors (Fig. 1).

192 **Fig 1.**

193

194 In the present study, the k-fold cross validation approach (Efron and Tibshirani, 1994) was
195 utilized to obtain training and testing datasets for each PTF. The number of folds (i. e., k) was
196 obtained by trial and error. To do so, some PTFs, selected randomly, were developed with 10, 15
197 and 20-fold cross-validation. Then, the k value which resulted in the best performance of the
198 PTFs, was selected to develop all PTFs in this study. The results showed that 20-fold cross
199 validation performed better than the other folds in most of the PTFs (Table 1). Therefore, 20-fold
200 cross validation was selected to develop PTFs in this study. Based on this approach, the 223
201 samples were randomly divided into 20 subsets and 20 models were developed by each
202 predicting technique for each PTF. In each model, training and testing datasets were based on a
203 ratio of 19:1. Finally, the average of the results of 20 models was calculated for each PTF.
204 Therefore, all data were used for the training and testing steps of the PTFs.

205 **Table 1-**

206 *2.5 Description of modeling techniques*

207 *2.5.1 Multiple nonlinear regression*

208 A NLR model based on a second-order polynomial for the prediction of the response variable y
209 from a number of p predictors can be written as (Rawls and Brakensiek, 1985):

$$y = a + \sum_{i=1}^p (b_i x_i + c_i x_i^2) \quad (2)$$

210 where a is the intercept, and two regression coefficients b_i and c_i are determined for every input
211 variable x_i .

212

213 *2.5.2 Random forest: an ensemble of regression trees*

214 RF has become a popular tool for regression and classification problems. The RF is an ensemble
215 method based on the regression tree methodology (i.e., CART) that was introduced for better

216 performance (Breiman, 2001). The model building process in the RF is the same as that in the
 217 CART method but without pruning (Breiman, 1984). Also, whereas a regression tree only grows
 218 by a single tree the RF grows by forest of trees. In other words, unlike a regression tree, in the
 219 RF for each tree only a subset of the input variables is applied. The number of inputs in each tree
 220 and also the number of trees in the forest can be distinct and it depends on the dataset. Least-
 221 squares boosting (LSBoost) fits regression ensembles. At every step, the ensemble fits a new
 222 learner to the difference between the observed response and the aggregated prediction of all
 223 learners grown previously. The ensemble fits to minimize the mean-squared error (MathWorks,
 224 2018). The number of trees used here was 16 which was established by trial and error. An
 225 architecture of the RF algorithm is shown in Fig. 2 where input matrix X consists of N samples
 226 and M input variables (sample set $S = [(x_i, y_i), i = 1, 2, \dots, N]$, $(X, Y) \in \mathbb{R}^{M \times R}$). The bootstrap
 227 method is utilized to construct n tree sample sets from the sample set S . At each bootstrap
 228 sample, about one-third of the dataset S was utilized as out of the bootstrap data or out-of-bag
 229 (*OOB*) data; whereas the rest is called in-bag data (Ibrahim and Khatib, 2017) (Fig. 2). Modeling
 230 of the regression tree is done for each sample set. In the RF algorithm, all individual trees give a
 231 predictive result. The final prediction value is calculated based on an average result of all
 232 individual trees (Wiesmeier et al., 2011). The prediction error is defined as follows (Liaw and
 233 Wiener, 2002):

$$MSE_{OOB} = \frac{\sum_{i=1}^{n_{tree}} (y_i - \hat{y}_i^{OOB})^2}{n_{tree}} \quad (3)$$

234 where MSE_{OOB} is the mean square error of the *OOB* data prediction, n_{tree} is the number of trees,
 235 and y_i and \hat{y}_i^{OOB} are the actual value of the *OOB* data and the average of all *OOB* predictions,
 236 respectively. Among all the ensemble methods, the RF method has high capability in solving

237 classification and regression problems, because the RF method combines several simple
 238 regression trees to better optimize prediction (Zaklouta and Stanculescu, 2012). The RF method
 239 increases differences for each single tree through random selection of the training samples and
 240 different variables at each splitting node. In the present study, the NLR and RF algorithms were
 241 implemented by fitlm and fitensemble functions in the MATLAB software, respectively.
 242 (MathWorks, 2018).

243 **Fig. 2.**

244

245 2.6 Evaluation criteria

246 The estimated water content was computed by estimated parameters of the van Genuchten model
 247 for each PTF at matric suctions from 0 to 1500 kPa. For curve-by-curve comparison of the
 248 measured and predicted SWRCs, different evaluation statistics were used. Various statistical
 249 criteria including integral root mean square error (*IRMSE*), integral mean error (*IME*) (Tietje and
 250 Tapkenhinrichs, 1993), Akaike's information criterion (*AIC*) (Akaike, 1974) and coefficient of
 251 determination (R^2) (Wösten et al., 2001), were utilized to assess the predictive ability of the RF
 252 and NLR algorithms, which are defined as:

$$IRMSE (cm^3 cm^{-3}) = \left[\frac{1}{b-a} \int_a^b (\hat{y}_i - y_i)^2 d \log|h| \right]^{\frac{1}{2}} \quad (4)$$

$$IME (cm^3 cm^{-3}) = \frac{1}{b-a} \int_a^b (\hat{y}_i - y_i) d \log|h| \quad (5)$$

$$AIC = N \times \ln \left[\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N} \right] + 2P \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2} \quad (7)$$

253

254 where h is the matric suction (kPa), y_i , \hat{y}_i and \bar{y}_i are the measured, predicted and average of
 255 the measured values of the water content, respectively, a and b values define the matric suction
 256 range over which the experimental curve is measured, i.e., 0 and 1500 kPa, respectively, and P
 257 and N are the number of parameters and the number of points that were considered in the SWRC,
 258 respectively. In calculating the AIC , N is the total number of points that were considered in the
 259 SWRC of all soil samples (i. e., $N = \text{number of soil samples} \times \text{number of paired points of the}$
 260 suction-water content for each soil sample), and i is paired points of the suctions-water content
 261 of the SWRC of each soil sample.

262 To evaluate the performance of each method in different PTFs, the effect of method as the first
 263 factor at two levels in the training step (i.e., NLR and RF methods) and at three levels in the
 264 testing step (i.e., NLR, RF and Rosetta methods), and the different PTFs as the second factor at
 265 15 levels (PTF1 to PTF15), were investigated using a two-way analysis of variance (ANOVA)
 266 with a randomized complete block design, based on the $IRMSE$ of prediction of the SWRC. The
 267 $IRMSE$ criterion calculates the total error, including bias and random errors, and is a more
 268 appropriate criterion for evaluating the accuracy and reliability of the RF and NLR methods
 269 compared to other criteria (Chai and Draxler, 2014). Therefore, to compare the predicting
 270 accuracy and reliability of the RF and NLR methods, the average values of the $IRMSE$ was
 271 compared with Duncan's test by MathWorks (2018) software.

272

273 3 Results and discussion

274 3.1 Descriptive statistics of the soil properties

275 Table 2 summarizes some basic descriptive statistics for soil variables of the entire dataset used
276 for the development of the PTFs. It can be seen that the average and maximum of clay content
277 were 21.4 and 48%, respectively. The OM ranged from 0.17 to 4.41% with a mean of 1.84%,
278 which was low due to the arid and semi-arid climates of Iran. The variation in soil texture is
279 shown graphically in the United States Department of Agriculture (USDA) textural triangle (Fig.
280 3). Considering the distribution and range of the variables (Fig. 3 and Table 2), the dataset can be
281 considered as representative of soils in arid and semi-arid regions of Iran.

282 **Table 2**

283 **Fig. 3.**

284 3.2 Correlation of input and output variables

285 The simple correlation coefficients between all variables are depicted by matrix plot in Fig. 4.
286 Correlation analysis was done between normalized input and output variables. The correlation test
287 was not performed for the θ_r variable, because its value was zero in 138 out of 223 soil samples,
288 as has been reported in other studies (Campbell and Horton Jr, 2002; Rawls et al., 1991; Tomasella
289 et al., 2000) for θ_r variable. Clay and sand contents, θ_{FC} , θ_{PWP} , d_g and OM had the greatest
290 significant correlations with the parameters of the van Genuchten model (Fig. 4), which was
291 consistent with other studies (Dexter et al., 2008; Nemes et al., 2006). For example, the correlation
292 coefficient between clay content and θ_s ($r = 0.323$) is close to that between OM and θ_s ($r = 0.268$).
293 Also, the results showed that there were significant correlations between θ_{PWP} and input variables
294 of clay content (+), sand content (-), BD (-), OM (+) and K_s (-), and also between θ_{PWP} and θ_s (+)
295 and n (-) parameters of the van Genuchten model (Fig. 4). Botula et al. (2012) also found the same

296 observation for the correlation of θ_{PWP} with sand and clay contents and BD of tropical Lower
 297 Congo soils. Nevertheless, with regard to these correlation coefficients, clay and sand contents,
 298 θ_{FC} , d_g and OM can be used for developing PTFs to estimate the SWRC. On the contrary, there
 299 was no correlation between K_s and the van Genuchten model parameters. There are many cases,
 300 where two variables might not show a strong simple correlation, but may show a strong association
 301 in the regression, along with other predictors. In other words, the simple correlation coefficient is
 302 a way to show the relationship between independent and dependent variables, but it cannot show
 303 a model for the relationship between these two variables, when other independent variables have
 304 been used in a multiple regression (Simmons et al., 2011). The result of multiple regression
 305 analysis with backward selection method showed that the K_s variable remained in the PTF14 and
 306 PTF15 for all the van Genuchten model parameters. Some of the regression equations with
 307 backward selection method are shown in the following as examples:

$$\theta_r = -0.69 + 0.22 \times \text{Clay} + 0.278 \times \text{Sand} + 0.20 \times K_s, R = 0.31^{**} \quad (8)$$

$$\alpha = -3.72 + 0.23 \times \text{Clay} + 0.17 \times \text{BD} + 0.282 \times K_s, R = 0.33^{**} \quad (9)$$

$$n = -1.76 + 0.24 \times \text{Sand} + 0.164 \times K_s, R = 0.30^{**} \quad (10)$$

308 On the other hand, the non-linear correlations between variables are very important in this study.
 309 Both the multiple NLR approach and RF data mining technique are non-linear prediction
 310 methods. Fig. 4 only shows simple linear correlation between variables, but there may be non-
 311 linear correlations between variables, which may affect the estimation of the dependent
 312 variables. For example, the results of non-linear correlations showed that K_s had strong
 313 correlations with θ_s and α of the van Genuchten model parameters by logarithmic ($\theta_s = 0.652 -$
 314 $0.027 \times \ln K_s, R = 0.62^{**}$) and power ($\alpha = 0.007 \times K_s^{0.283}, R = 0.57^{**}$) equations, respectively, which
 315 were greater than their simple correlations

316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339

Fig. 4.

3.3 Development of the PTFs using the RF and NLR methods

Results of the multicollinearity analysis (*VIF*) are shown in Table 3. The *VIF* values showed low levels of multicollinearity among the independent variables ($VIF < 10$) (Khodaverdiloo et al., 2011).

Table 3-

3.3.1 Comparing the accuracy and reliability of the RF and NLR methods

Table 4 shows the results of the ANOVA of the *IRMSE* of prediction of the SWRC by different methods and PTFs. The effect of methods and PTFs, and their interaction, on the *IRMSE* was significant at $P < 0.01$, 0.01 and 0.05, respectively, in the training step, and at $P < 0.01$, 0.01 and 0.01, respectively, in the testing step. Therefore, we focus on the results and discussion of the comparison of the method \times PTF interaction effects.

Table 4-

Results of the prediction of the SWRC through the van Genuchten model using the NLR and RF-based PTFs are depicted in Figs. 5 and 6 for the training and testing steps, respectively. The accuracy and reliability are used to express the performance of the PTFs in the training and testing steps, respectively.

The results of the first to fourth steps of the training dataset (Fig. 5) showed that the RF method had better performance compared to the NLR method for the prediction of the SWRC in all PTFs in terms of the *IRMSE* and R^2 criteria and the differences were significant ($P < 0.05$) for PTFs 2, 3, 6, 7, 10, 13, 14 and 15 in terms of the *IRMSE* criterion. Also, the accuracy of the RF method was better than that of the NLR method in 80% of the PTFs (with the exception of the PTFs 5, 9 and 12) in terms of the *AIC* criterion. In the training step, the values of the *IRMSE* of the first to

340 fourth steps for the NLR model varied from 0.030 to 0.063 cm³ cm⁻³ and these were larger than
341 those in the RF model, which ranged from 0.028 to 0.061 cm³ cm⁻³, respectively. Also, the
342 values of the R^2 of the first to fourth steps for the RF model varied from 0.981 to 0.992, and this
343 was larger than those in the NLR model, which ranged from 0.979 to 0.991 (Fig. 5).

344 The results of the first to fourth steps of the testing dataset (Fig. 6) showed that the NLR method
345 had a better performance compared to the RF method on the prediction of the SWRC for PTFs 5,
346 8, 9 and 15 only in terms of the *IRMSE* criterion (significant at $P<0.05$). In the other PTFs there
347 were no significant differences between the *IRMSE* of the two methods and the R^2 and *AIC*
348 criteria were comparable. In the testing step, the values of the *IRMSE* and *AIC* of the first to
349 fourth steps for the RF models varied from 0.038 to 0.065 cm³ cm⁻³ and from -13476.2 to -
350 17646.8, respectively, and these were comparable to those of the NLR models (with the
351 exception of PTF1), which ranged from 0.032 to 0.064 cm³ cm⁻³ and from -14096.1 to -19234.1,
352 respectively (Fig. 6). Also, the values of the R^2 of the first to fourth steps for the NLR models
353 varied from 0.979 to 0.989, and this was comparable to those of the RF models for all PTFs,
354 which ranged from 0.977 to 0.987 (Fig. 6).

355 In each of the PTFs 1 to 5, the NLR and RF methods performed better ($P<0.05$) than the Rosetta
356 PTFs. Fig. 6(A) shows that the Rosetta-based PTFs had greater values of the *IME* criterion
357 compared to the NLR and RF-based PTFs. The reason can be attributed to the various methods
358 of optimizing parameters. The Rosetta method has only one ANN type with particular structure.
359 In other words, the number of hidden layers (one) and neurons (six) and also the activation
360 function (tangent hyperbolic) are constant for prediction of the SWRC in the Rosetta software.
361 Therefore, the Rosetta method is not a dynamic approach for optimization, whereas the
362 parameters of the RF method, such as number of splits and trees, and learning rate continuously

363 and dynamically, change to achieve the best result of the objective function. The Rosetta method
364 was developed from a large dataset, while the soils used in the present study were collected from
365 a completely different climate area that was not represented in the Rosetta's database. Also,
366 presented RF and NLR models were trained using this particular dataset while Rosetta had been
367 trained using a different dataset. In other words, the results of the PTFs in the testing step were
368 based on a soil dataset used for training. This could be a reason for Rosetta's poor performance
369 compared with the RF and NLR methods. As a result, it seems that the universal portability of
370 the Rosetta method can be limited. The testing results are in agreement with Touil et al. (2016)
371 who found that the parametric-based PTFs of nonlinear models gave a better prediction than the
372 Rosetta PTFs. The Figs. 5(A) and 6(A) showed that all of the *IME* values were negative for all
373 PTFs at the training and testing steps. There are regular errors (bias) in the prediction of the
374 SWRC that can be corrected by finding a correction coefficient, which would improve the
375 accuracy and reliability of the estimations (Bayat et al., 2015).

376 **Fig. 5.**

377 **Fig. 6.**

378
379 The RF method in the training section gave better predictions of the SWRC compared to the
380 NLR method (Fig. 5). The RF method produces low bias and variation in the data by majority
381 voting compared to a single regression tree (Cheng et al., 2019; Matin and Chelgani, 2016). In
382 this connection, the results of the standard deviations (SD) of evaluation criteria in each PTF for
383 the training step (Fig. 5) showed that the RF method had a lower variation than the NLR method.
384 Accordingly, the values of SD for the *IRMSE* and R^2 criteria were 0.024 and 0.022, respectively,
385 for the NLR model and these were larger than those in the RF model, which were 0.020 and

386 0.017, respectively, for the training step. On the other hand, the RF method can be applied to
387 high dimensional datasets in regressions (Janitza et al., 2016; Zhao et al., 2016).

388 As depicted in Fig. 6, unlike in the training section, the NLR method gave better predictions in
389 the testing section compared to the RF method for the prediction of the SWRC. In other words,
390 the reliability of the NLR method was better than that of the RF method in all the PTFs. The
391 NLR equations can be more useful than the MLR method for the prediction of the SWRC due to
392 their high flexibility (Williams et al., 1992). In other words, the NLR models have capacity to
393 capture nonlinear relationships in the dataset. Tomasella et al. (2000) successfully developed
394 parametric PTFs for soils of the humid tropics using polynomials of n^{th} order. Medrado and Lima
395 (2014) successfully developed NLR-based PTFs to predict the four parameters of the van
396 Genuchten model for Brazilian soils. Also, Touil et al. (2016) developed parametric-PTFs to
397 predict the SWRC using the NLR method from more readily-available properties such as soil
398 texture, OM content, and BD for 242 soil samples of Algeria. They reported that the parametric-
399 PTFs had better performance than Rosetta-based PTFs.

400 In the present study, in contrast to the NLR method which had less differences between the error
401 values of the training and testing steps, the error values of the RF method in the testing dataset
402 were much greater than those in the training dataset. These results can be due to overprediction
403 phenomenon in the RF method. Gupta et al. (2017) expressed that one of the disadvantages of
404 the RF method is the overprediction. In other words, the RF method is a ‘greedy’ method that
405 easily leads to overprediction and instability in the testing step and solving this problem can be
406 of great significance for improving the reliability of the RF method (Liu, 2014). Also, Ma et al.
407 (2005) reported instability in results of the RF method. The forest size developed by the RF has
408 not been clearly defined (Liu, 2014). Therefore, oversized scale can decrease the reliability and

409 efficiency of the SWRC prediction. Hong et al. (2016) evaluated landslide susceptibility maps
410 produced using the RF method and compared these maps with those from statistical-based
411 methods, such as logistic regression, and their study revealed that the performance of the
412 statistical-based methods was better than that of the RF method. A similar result was reported by
413 Esposito et al. (2014). Generally, RFs are best suited for problems with many input variables and
414 a reasonable sample size. According to our results (Figs. 5 and 6), performance of the PTFs was
415 improved by increasing the number of input variables.

416 3.3.2 Evaluation of the effect of the basic soil properties on prediction performance of the 417 SWRC

418 A significant improvement was achieved in the accuracy of PTF5 (with the inputs of Sand
419 content+Clay content+BD+ θ_{FC} + θ_{PWP}) compared to other PTFs (with the exception of PTFs 4, 8,
420 9, 11 and 12) by both NLR and RF methods in terms of the *IRMSE* criterion (Fig. 5). Among the
421 PTFs of each method (RF or NLR), PTF5 had the greatest R^2 (0.992 and 0.991, respectively) and
422 the smallest *IRMSE* (0.028 and 0.03, respectively) and *AIC* (-19432 and -19571.1, respectively)
423 in the training step of the prediction of the SWRC. In connection with the importance of input
424 variables, an improvement was achieved in the reliability of the prediction of the SWRC by PTFs
425 9 (with the inputs of d_g + δ_g +BD+ θ_{FC} + θ_{PWP}) and 12 (with the inputs of Sand content+Clay
426 content+TP+ θ_{FC} + θ_{PWP}) from the second and third steps, using the NLR (*IRMSE*=0.032 cm³ cm⁻³
427 ³, *AIC*=-19234.1 and R^2 =0.989) and RF (*IRMSE*=0.038 cm³ cm⁻³, *AIC*=-17646.8 and R^2 =0.987)
428 methods, respectively, in comparison with the other PTFs of each method (Fig 6). However, the
429 differences of PTFs 9 and 12 were not significant ($P<0.05$) with PTFs 4, 5, 8, 11 and 12 in the
430 NLR method and with PTFs 4, 5, 8, 9 and 11 in the RF method, respectively, in terms of the
431 *IRMSE* criterion.

432

433 3.3.2.1 *Effect of using different input variables of PSD and soil structure as predictors on the*
434 *SWRC prediction*

435 To evaluate the effect of using different descriptors of the PSD on the prediction of the SWRC,
436 PTFs 2, 3, 4 and 5 (clay and sand contents) from the first step were compared with PTFs 6, 7, 8
437 and 9 (d_g and δ_g) from the second step, respectively. In the same way, to evaluate the effect of
438 using different descriptors of soil structure on the prediction of the SWRC, PTFs 3, 4 and 5 (BD)
439 were compared with PTFs 10, 11 and 12 (TP) from the third step, respectively. The accuracy and
440 reliability of the prediction of the SWRC by both NLR and RF methods were not significantly
441 different ($P < 0.05$) (Figs. 5B and 6B). For descriptors of soil structure, the accuracy and
442 reliability of the prediction of the SWRC by both NLR and RF methods decreased in terms of the
443 *IRMSE* criterion for PTFs 10 to 12 from the third step compared to PTFs 3 to 5 (with the
444 exception of PTFs 11 and 12 in the testing step for the RF method), respectively, when TP was
445 used instead of BD in the list of input variables (Figs. 5B and 6B). However, the differences
446 were not significant ($P < 0.05$).

447 The lack of significant differences between textural contents (clay and sand contents) and
448 statistics (d_g and δ_g), and also between TP and BD on the SWRC prediction can be due to
449 correlation of these parameters with the parameters of the van Genuchten model (Fig. 4). The
450 SWRC is strongly influenced by the soil structure or pore-size distribution and soil texture at
451 small and great matric suctions, respectively (Pachepsky et al., 2006). Therefore, input variables
452 of the textural contents or statistics can influence the residual saturation region of the SWRC.
453 However, soil water content at the dry end (high matric suctions) of the SWRC is primarily
454 determined by textural contents (Hillel, 1998). Also, TP and BD are indicators of soil structure

455 and had significant correlations with θ_s (Fig. 4). Indeed, TP was calculated by BD and particle
456 density (Rab et al., 2011). The d_g and δ_g predictors were derived from soil textural contents
457 (Shirazi and Boersma, 1984). Therefore, these could be reasons for similar effects of textural
458 contents and statistics and also TP and BD predictors on the prediction of the SWRC.
459 Many researchers used textural contents (Adhikary et al., 2008; Chakraborty et al., 2011;
460 Minasny et al., 1999; Tomasella and Hodnett, 1998), d_g and δ_g (Rab et al., 2011; Scheinost et al.,
461 1997; Ungaro et al., 2005), BD (Bayat et al., 2011; Pachepsky et al., 1998) and TP (Bayat et al.,
462 2011; Pachepsky et al., 1998; Schaap et al., 1998) as effective predictors to derive point- and
463 parametric-PTFs. Nemes et al. (2003), Schaap et al. (2001) and Schaap et al. (1998) reported that
464 the variables of PTF5 have better capability on predicting the parameters of the van Genuchten
465 (1980) model with an average *RMSE* of 0.026, 0.044 and 0.058 $\text{cm}^3\text{cm}^{-3}$, respectively.
466 According to the results of the accuracy (Fig. 5) and reliability (Fig. 6) of PTFs 5, 9 and 12, it
467 seems that certain points of the SWRC (e.g., θ_{FC}) can help to improve the prediction of the
468 SWRC and this is in agreement with Schaap et al. (2001). These results indicate that the presence
469 of at least one moisture point (e.g., θ_{FC}) can improve the prediction of the SWRC. In the first
470 step, PTF5 with two moisture points ($\theta_{FC}+\theta_{PWP}$) and PTF4 with one moisture point (θ_{FC})
471 improved the prediction of the SWRC by 55, 48, 42% and 51, 44, 38% in terms of the *IRMSE*
472 criterion compared to the PTFs 1, 2 and 3, respectively, in the RF method in the training step. In
473 the testing section of the second step, PTF9 with two moisture points ($\theta_{FC}+\theta_{PWP}$) and PTF8 with
474 one moisture point (θ_{FC}) decreased the *IRMSE* by 49, 44% and 44, 39% compared to PTFs 6 and
475 7, respectively, in the NLR method. The points above are also true for the RF-based PTF12 in
476 the third step of the testing section. Many researchers successfully applied θ_{FC} and θ_{PWP} as

477 effective predictors to derive point- and parametric-PTFs (Børgesen and Schaap, 2005; Nemes et
478 al., 2003; Schaap et al., 2001; Touil et al., 2016; Twarakavi et al., 2009).

479

480 3.3.2.2 *Effect of using OM and K_s as predictors on the SWRC prediction*

481 To evaluate the effect of using OM and/or K_s and points of the SWRC on the prediction of the
482 SWRC, the performances of PTFs 13, 14 and 15 were compared with those of PTFs 4 and 5. The
483 accuracy and reliability of the prediction of the SWRC by both NLR and RF methods,
484 significantly ($P<0.05$) decreased in terms of the *IRMSE*, for the PTFs 13, 14 and 15 from the
485 fourth step, when OM and/or K_s were used with textural contents and BD as inputs instead of θ_{FC}
486 or both θ_{FC} and θ_{PWP} in the list of input variables, compared to PTFs 4 and 5 at the first step
487 (Figs. 5B and 6B). Therefore OM and K_s were not as effective predictors as θ_{FC} and θ_{PWP} in the
488 prediction of the SWRC, because θ_{FC} and θ_{PWP} are two points of the SWRC and enter direct
489 information of the SWRC into the PTFs, whereas OM and K_s enter indirect information, and
490 therefore had less effect in the improvement of the estimation of the SWRC. These results agreed
491 well with results obtained by Børgesen and Schaap (2005). They reported that PTFs with the
492 inputs of θ_{FC} and θ_{PWP} had smaller *RMSE* values than a PTF with the input of OM (0.038 versus
493 0.042) in the prediction of the SWRC. On the other hand, the results showed that by adding OM
494 and/or K_s as predictors in the PTFs 13, 14 and 15, the accuracy (Fig. 5B) and reliability (Fig. 6B)
495 of the prediction of the SWRC improved by 16, 13, 17 and 7.1, 6.3, 6.9%, respectively,
496 compared to the PTF3 in terms of the *IRMSE* criterion in the RF method.

497 The SWRC depends mainly on the soil texture and structure (Hillel, 1998), with OM affecting
498 the SWRC through development of soil structure (Nemes et al., 2005), important at low suctions.
499 However, the OM retains water itself. Similarly, K_s can be a descriptive index of soil texture and

500 porosity (Hillel, 1998). The correlation results showed that K_s can be strongly influenced by clay
501 content and textural statistics (d_g and δ_g) (Fig. 4). Bayat et al. (2013b) applied OM and K_s to
502 estimate water content at the measured matric suctions. They found that the OM and K_s can be
503 most appropriately used in point-based PTFs to estimate water content at the matric suctions of
504 25 and 50 kPa. Also, the result of the present study agreed well with results obtained by Hollis et
505 al. (1977) and Rawls et al. (1983). In this study, the OM and K_s in the PTFs 13, 14 and 15 were
506 not effective predictors compared to θ_{FC} and θ_{PWP} in the PTFs 4 and 5, otherwise they had better
507 results than PTF3.

508

509 **4 Conclusion**

510 Machine-learning tools have been widely applied for the prediction of the SWRC. The present
511 study evaluated the capability and performance of the RF method as a novel machine learning
512 tool and compared its performance with that of the NLR method on the prediction of the SWRC,
513 using different combinations of easily-available soil properties. It was found that the RF method
514 had a better performance ($P < 0.05$) than the NLR method in the training step of the prediction of
515 the SWRC in term of the *IRMSE*, *AIC* and R^2 criteria. However, in the testing step, NLR had a
516 better performance than RF. The poor performance of the RF compared to the NLR method
517 could be due to overprediction in the former, resulting in instability in the testing step. The RF
518 method can be sensitive to sparse areas on the prediction space. In other words, the performance
519 and sensitivity of predictions, and the computational intensity of the RF method depends on the
520 distribution and number of observations and input variables. Therefore, the method should be
521 tested further with different datasets to evaluate its performance through soil and water
522 investigations. An improvement was achieved in the accuracy of the prediction of the SWRC in

523 the training step of the PTF5 (with the inputs of Sand content+Clay content+BD+ θ_{FC} + θ_{PWP}) by
524 both NLR and RF methods and also an improvement was achieved in the reliability of the PTF9
525 (with the inputs of $d_g+\delta_g$ +BD+ θ_{FC} + θ_{PWP}) and PTF12 (with the inputs of Sand content +Clay
526 content+TP+ θ_{FC} + θ_{PWP}) by the NLR and RF methods compared to other PTFs, respectively.
527 Considering that the PTFs 5, 9, and 12 had no significant difference from PTF4 (with the inputs
528 of Sand content+Clay content+BD+ θ_{FC}) and PTF8 (with the inputs of $d_g+\delta_g$ +BD+ θ_{FC} + θ_{PWP}),
529 these latter PTFs, with less and more-easily measured input variables, are suggested to be the
530 best PTFs for the prediction of the SWRC. Also, PTFs without predictors of θ_{FC} and θ_{PWP} , such
531 as the PTF3 (with the inputs of Sand content+Clay content+BD) and PTF7 (with the inputs of
532 $d_g+ \delta_g$ +BD), can be effective models for the prediction of the SWRC.

533

534 **Acknowledgements**

535 This work was funded by Bu Ali Sina University, Hamedan, Iran. The authors are deeply
536 grateful to anonymous reviewers and the editor for their helpful comments on the manuscript.

537

538 **References**

539 Adhikary, P.P., Chakraborty, D., Kalra, N., Sachdev, C., Patra, A., Kumar, S., Tomar, R.,
540 Chandna, P., Raghav, D., Agrawal, K., 2008. Pedotransfer functions for predicting the
541 hydraulic properties of Indian soils. *Soil Res.* 46, 476-484.
542 Akaike, H., 1974. A new look at the statistical model identification. *IEEE transactions on*
543 *automatic control* 19, 716-723.

544 Araya, S.N., Ghezzehei, T.A., 2019. Using Machine Learning for Prediction of Saturated
545 Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. *Water Resour.*
546 *Res.* 55, 5715-5737.

547 Bayat, H., Ersahin, S., Hepper, E.N., 2013a. Improving estimation of specific surface area by
548 artificial neural network ensembles using fractal and particle size distribution curve
549 parameters as predictors. *Environ. Model Assess.* 18, 605-614.

550 Bayat, H., Neyshabouri, M., Mohammadi, K., Nariman-Zadeh, N., 2011. Estimating water
551 retention with pedotransfer functions using multi-objective group method of data
552 handling and ANNs. *Pedosphere* 21, 107-114.

553 Bayat, H., Neyshaburi, M.R., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2013b.
554 Improving water content estimations using penetration resistance and principal
555 component analysis. *Soil Tillage Res.* 129, 83-92.

556 Bayat, H., Sedaghat, A., Sinegani, A.A.S., Gregory, A.S., 2015. Investigating the relationship
557 between unsaturated hydraulic conductivity curve and confined compression curve. *J.*
558 *Hydrol.* 522, 353-368.

559 Berry, W.D., 1993. *Understanding regression assumptions.* Sage Publications, London.

560 Blake, G., Hartge, K., 1986. Bulk density, *Methods of Soil Analysis: Part 1. Physical and*
561 *Mineralogical Methods,* Madison, Wisconsin, USA: Soil Sci. Soc. Am. J.

562 Børgesen, C.D., Schaap, M.G., 2005. Point and parameter pedotransfer functions for water
563 retention predictions for Danish soils. *Geoderma* 127, 154-167.

564 Botula, Y.-D., Cornelis, W., Baert, G., Van Ranst, E., 2012. Evaluation of pedotransfer functions
565 for predicting water retention of soils in Lower Congo (DR Congo). *Agric. Water Manag.*
566 111, 1-10.

567 Botula, Y.-D., Cornelis, W.M., Baert, G., Mafuka, P., Van Ranst, E., 2013. Particle size
568 distribution models for soils of the humid tropics. *Journal of Soils and Sediments* 13,
569 686-698.

570 Bouma, J., 1989. Using soil survey data for quantitative land evaluation, *Advances in soil*
571 *science*. Springer, pp. 177-213.

572 Breiman, L., 1984. *Classification and regression trees*. Routledge, New York.

573 Breiman, L., 2001. Random forests. *Machine learning* 45, 5-32.

574 Bruce, R.R., Luxmoore, R.J., 1986. Water Retention: Field Methods, In: Klute, A. (Ed.),
575 *Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods*. Soil Science
576 *Society of America, American Society of Agronomy, Madison, WI, pp. 663-686.*

577 Campbell, G.S., Horton Jr, R., 2002. *Methods of Soil Analysis: Part 4, Physical Methods*. Soil
578 *Sci. Soc. Am.*

579 Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?—
580 *Arguments against avoiding RMSE in the literature*. *Geosci. Model Dev.* 7, 1247-1250.

581 Chakraborty, D., Mazumdar, S., Garg, R., Banerjee, S., Santra, P., Singh, R., Tomar, R., 2011.
582 *Pedotransfer functions for predicting points on the moisture retention curve of Indian*
583 *soils*. *Indian J. Agr. Sci.* 81, 1030.

584 Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method
585 *approach to model travel mode choice behavior*. *Travel behaviour and society* 14, 1-10.

586 Dexter, A., Czyż, E., Richard, G., Reszkowska, A., 2008. A user-friendly water retention
587 *function that takes account of the textural and structural pore spaces in soil*. *Geoderma*
588 143, 243-253.

589 Dobarco, M.R., Cousin, I., Le Bas, C., Martin, M.P., 2019. Pedotransfer functions for predicting
590 available water capacity in French soils, their applicability domain and associated
591 uncertainty. *Geoderma* 336, 81-95.

592 Efron, B., Tibshirani, R.J., 1994. *An introduction to the bootstrap*. CRC press.

593 Esposito, C., Barra, A., Evans, S.G., Scarascia Mugnozza, G., Delaney, K., 2014. Landslide
594 susceptibility analysis by the comparison and integration of random forest and logistic
595 regression methods; application to the disaster of Nova Friburgo-Rio de Janeiro, Brasil
596 (January 2011), EGU General Assembly Conference Abstracts.

597 Gee, G.W., Or, D., 2002. 2.4 Particle-Size Analysis, In: Dane, J.H., Topp, C.G. (Eds.), *Methods*
598 *of Soil Analysis: Part 4 Physical Methods*. Soil Science Society of America, Madison,
599 WI, pp. 255-293.

600 Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., 2019a. Machine Learning
601 Approaches to Develop Pedotransfer Functions for Tropical Sri Lankan Soils. *Water* 11,
602 1940.

603 Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., Amarasekara, M., 2019b.
604 Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. *Soil*
605 *Till. Res.* 190, 109-119.

606 Gupta, B., Rawat, A., Jain, A., Arora, A., Dharmi, N., 2017. Analysis of various decision tree
607 algorithms for classification in data mining. *Int. J. Comput. Appl.* 163, 15-19.

608 Haghverdi, A., Leib, B.G., Cornelis, W.M., 2015. A simple nearest-neighbor technique to predict
609 the soil water retention curve. *Transactions of the ASABE* 58, 697-705.

610 Hillel, D., 1998. *Environmental soil physics: Fundamentals, applications, and environmental*
611 *considerations*. Academic press.

612 Hocking, R.R., 2013. Methods and applications of linear models: regression and the analysis of
613 variance. John Wiley & Sons.

614 Hollis, J., Jones, R., Palmer, R., 1977. The effects of organic matter and particle size on the
615 water-retention properties of some soils in the West Midlands of England. *Geoderma* 17,
616 225-238.

617 Hong, H., Pourghasemi, H.R., Pourtaghi, Z.S., 2016. Landslide susceptibility assessment in
618 Lianhua County (China): a comparison between a random forest data mining technique
619 and bivariate and multivariate statistical models. *Geomorphology* 259, 105-118.

620 IBM, C., 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

621 Ibrahim, I.A., Khatib, T., 2017. A novel hybrid model for hourly global solar radiation prediction
622 using random forests technique and firefly algorithm. *Energy Convers. Manag.* 138, 413-
623 425.

624 Janitza, S., Tutz, G., Boulesteix, A.-L., 2016. Random forest for ordinal responses: prediction
625 and variable selection. *Comput. Statist. Data Anal.* 96, 57-73.

626 Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced
627 pedotransfer functions with support vector machines to predict water retention of
628 calcareous soil. *Eur. J. Soil Sci.* 67, 276-284.

629 Khodaverdiloo, H., Homaei, M., van Genuchten, M.T., Dashtaki, S.G., 2011. Deriving and
630 validating pedotransfer functions for some calcareous soils. *J. Hydrol.* 399, 93-99.

631 Klute, A., 1986. Water Retention: Laboratory Methods, In: Klute, A. (Ed.), *Methods of Soil*
632 *Analysis: Part 1—Physical and Mineralogical Methods*. Soil Science Society of America,
633 American Society of Agronomy, Madison, WI, pp. 635-662.

634 Klute, A., Dirksen, C., 1986. Hydraulic Conductivity and Diffusivity: Laboratory Methods, In:
635 Klute, A. (Ed.), *Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods*.
636 Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 687-
637 734.

638 Koekkoek, E., Booltink, H., 1999. Neural network models to predict soil water retention. *Eur. J.*
639 *Soil Sci.* 50, 489-495.

640 Lamorski, K., Pachepsky, Y., Sławiński, C., Walczak, R., 2008. Using support vector machines
641 to develop pedotransfer functions for water retention of soils in Poland. *Soil Sci. Soc.*
642 *Am. J.* 72, 1243-1247.

643 Lamorski, K., Sławiński, C., Moreno, F., Barna, G., Skierucha, W., Arrue, J.L., 2014. Modelling
644 soil water retention using support vector machines with genetic algorithm optimisation.
645 *Sci. World J.* 2014, 740521, 1-10.

646 Liaw, A., Wiener, M., 2002. Classification and regression by random forest. *R news* 2, 18-22.

647 Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture:
648 comparison of regression tree and Random Forest models. *Geoderma* 170, 70-79.

649 Liu, Y., 2014. Random forest algorithm in big data environment. *Comput. Model. New Tech.* 18,
650 147-151.

651 Ma, Y., Cukic, B., Singh, H., 2005. A classification approach to multi-biometric score fusion,
652 *International Conference on Audio-and Video-Based Biometric Person Authentication*.
653 Springer, pp. 484-493.

654 MathWorks, 2018. *MATLAB: the language of technical computing*, Inc., Natick, Massachusetts,
655 United States.

656 Matin, S., Chelgani, S.C., 2016. Estimation of coal gross calorific value based on various
657 analyses by random forest method. *Fuel* 177, 274-278.

658 Medrado, E., Lima, J.E., 2014. Development of pedotransfer functions for estimating water
659 retention curve for tropical soils of the Brazilian savanna. *Geoderma Regional* 1, 59-66.

660 Merdun, H., Çınar, Ö., Meral, R., Apan, M., 2006. Comparison of artificial neural network and
661 regression pedotransfer functions for prediction of soil water retention and saturated
662 hydraulic conductivity. *Soil Tillage Res.* 90, 108-116.

663 Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the
664 development of pedotransfer functions for water-retention curves. *Geoderma* 93, 225-
665 253.

666 Mualem, Y., 1976. A new model for predicting the hydraulic conductivity of unsaturated porous
667 media. *Water Resour. Res.* 12, 513-522.

668 Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2005. Influence of organic matter on the estimation of
669 saturated hydraulic conductivity. *Soil Sci. Soc. Am. J.* 69, 1330-1337.

670 Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor
671 approach to estimate soil hydraulic properties. *Soil Sci. Soc. Am. J.* 70, 327-336.

672 Nemes, A., Schaap, M., Wösten, J., 2003. Functional evaluation of pedotransfer functions
673 derived from different scales of data collection. *Soil Sci. Soc. Am. J.* 67, 1093-1102.

674 Neyshaburi, M.R., Bayat, H., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2015.
675 Improvement in estimation of soil water retention using fractal parameters and
676 multiobjective group method of data handling. *Arch. Agron. Soil Sci.* 61, 257-273.

677 Nguyen, P.M., Haghverdi, A., De Pue, J., Botula, Y.-D., Le, K.V., Waegeman, W., Cornelis,
678 W.M., 2017. Comparison of statistical regression and data-mining techniques in
679 estimating soil water retention of tropical delta soils. *Biosyst. Eng.* 153, 12-27.

680 Pachepsky, Y., Rawls, W., Gimenez, D., Watt, J., 1998. Use of soil penetration resistance and
681 group method of data handling to improve soil water retention estimates. *Soil Tillage*
682 *Res.* 49, 117-126.

683 Pachepsky, Y.A., Rawls, W., 1999. Accuracy and reliability of pedotransfer functions as affected
684 by grouping soils. *Soil Sci. Soc. Am. J.* 63, 1748-1757.

685 Pachepsky, Y.A., Rawls, W., Lin, H., 2006. Hydropedology and pedotransfer functions.
686 *Geoderma* 131, 308-316.

687 Pachepsky, Y.A., Timlin, D., Varallyay, G., 1996. Artificial neural networks to estimate soil
688 water retention from easily measurable data. *Soil Sci. Soc. Am. J.* 60, 727-733.

689 Rab, M., Chandra, S., Fisher, P., Robinson, N., Kitching, M., Aumann, C., Imhof, M., 2011.
690 Modelling and prediction of soil water contents at field capacity and permanent wilting
691 point of dryland cropping soils. *Soil Res.* 49, 389-407.

692 Rajkai, K., Kabos, S., Van Genuchten, M.T., 2004. Estimating the water retention curve from
693 soil properties: comparison of linear, nonlinear and concomitant variable methods. *Soil*
694 *Tillage Res.* 79, 145-152.

695 Rawls, W., Brakensiek, D., Soni, B., 1983. Agricultural management effects on soil water
696 processes part I: Soil water retention and Green and Ampt infiltration parameters.
697 *Transactions of the ASAE* 26, 1747-1752.

698 Rawls, W., Gish, T., Brakensiek, D., 1991. Estimating soil water retention from soil physical
699 properties and characteristics, *Advances in soil science*. Springer, pp. 213-234.

700 Rawls, W.J., Brakensiek, D., 1985. Prediction of soil water properties for hydrologic modeling,
701 Watershed management in the eighties. ASCE, pp. 293-299.

702 Schaap, M.G., Leij, F.J., van Genuchten, M.T., 1998. Neural network analysis for hierarchical
703 prediction of soil hydraulic properties. *Soil Sci. Soc. Am. J.* 62, 847-855.

704 Schaap, M.G., Leij, F.J., van Genuchten, M.T., 2001. Rosetta: A computer program for
705 estimating soil hydraulic parameters with hierarchical pedotransfer functions. *J. Hydrol.*
706 251, 163-176.

707 Scheinost, A., Sinowski, W., Auerswald, K., 1997. Regionalization of soil water retention curves
708 in a highly variable soilscape, I. Developing a new pedotransfer function. *Geoderma* 78,
709 129-143.

710 Seo, S., 2006. A review and comparison of methods for detecting outliers in univariate data sets,
711 Thesis for Master of Science in Field of Public Health University of Pittsburgh, pp. 1-59.

712 Shirazi, M.A., Boersma, L., 1984. A unifying quantitative analysis of soil texture. *Soil Sci. Soc.*
713 *Am. J.* 48, 142-147.

714 Shwetha, P., Varija, K., 2015. Soil water retention curve from saturated hydraulic conductivity
715 for sandy loam and loamy sand textured soils. *Aquat. Procedia* 4, 1142-1149.

716 Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed
717 flexibility in data collection and analysis allows presenting anything as significant.
718 *Psychol. Sci.* 22, 1359-1366.

719 Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019.
720 Mapping soil hydraulic properties using random forest based pedotransfer functions and
721 geostatistics. *Hydrol. Earth Syst. Sci.* 23, 2615-2635.

- 722 Tietje, O., Tapkenhinrichs, M., 1993. Evaluation of pedo-transfer functions. *Soil Sci. Soc. Am. J.*
723 57, 1088-1095.
- 724 Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited
725 data in Brazilian Amazonia. *Soil Sci.* 163, 190-202.
- 726 Tomasella, J., Hodnett, M.G., Rossato, L., 2000. Pedotransfer functions for the estimation of soil
727 water retention in Brazilian soils. *Soil Sci. Soc. Am. J.* 64, 327-338.
- 728 Tóth, B., Makó, A., Toth, G., 2014. Role of soil properties in water retention characteristics of
729 main Hungarian soil types. *J. Cent. Eur. Agric.* 15, 137-153.
- 730 Touil, S., Degre, A., Chabaca, M.N., 2016. Sensitivity analysis of point and parametric
731 pedotransfer functions for estimating water retention of soils in Algeria. *Soil* 2, 647.
- 732 Twarakavi, N.K., Šimůnek, J., Schaap, M., 2009. Development of pedotransfer functions for
733 estimation of soil hydraulic parameters using support vector machines. *Soil Sci. Soc. Am.*
734 *J.* 73, 1443-1452.
- 735 Ungaro, F., Calzolari, C., Busoni, E., 2005. Development of pedotransfer functions using a group
736 method of data handling for the soil of the Pianura Padano–Veneta region of North Italy:
737 water retention properties. *Geoderma* 124, 293-317.
- 738 van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of
739 unsaturated soils. *Soil Sci. Soc. Am. J.* 44, 892-898.
- 740 Verhagen, J., 1997. Site specific fertiliser application for potato production and effects on N-
741 leaching using dynamic simulation modelling. *Agric. Ecosyst. Environ.* 66, 165-175.
- 742 Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: A survey
743 and results of new tests. *Pattern Recognit.* 44, 330-349.

744 Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil
745 organic matter, and a proposed modification of the chromic acid titration method. *Soil*
746 *Sci.* 37, 29-38.

747 Wassar, F., Gandolfi, C., Rienzner, M., Chiaradia, E.A., Bernardoni, E., 2016. Predicted and
748 measured soil retention curve parameters in Lombardy region north of Italy. *International*
749 *Soil and Water Conservation Research* 4, 207-214.

750 Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic
751 matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant Soil*
752 340, 7-24.

753 Williams, J., Ross, P., Bristow, K.L., 1992. Prediction of the Campbell water retention function
754 from texture, structure, and organic matter. In 'Indirect methods for estimating the
755 hydraulic properties of unsaturated soils.' University of California: Riverside.

756 Wösten, J., Pachepsky, Y.A., Rawls, W., 2001. Pedotransfer functions: bridging the gap between
757 available basic soil data and missing soil hydraulic characteristics. *J. Hydrol.* 251, 123-
758 150.

759 Zaklouta, F., Stanciulescu, B., 2012. Real-time traffic-sign recognition using tree classifiers.
760 *IEEE Transactions on Intelligent Transportation Systems* 13, 1507-1514.

761 Zhao, P., Su, X., Ge, T., Fan, J., 2016. Propensity score and proximity matching using random
762 forest. *Contemp. Clin. Trials* 47, 85-92.

763

764 **Figure captions**

765 **Fig. 1.** Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten
766 model parameters (θ_r , θ_s , α and n) of the soil water retention curve (SWRC). A list of
767 abbreviations is available in the notation box.

768 **Fig. 2.** An architecture of a random forest.

769 **Fig. 3.** Variation of soil texture classes for the dataset ($n = 223$) on the United States Department
770 of Agriculture (USDA) textural triangle.

771 **Fig. 4.** Correlation matrix plot between input and output variables.

772 ** Correlation is significant at the $P < 0.01$ level.

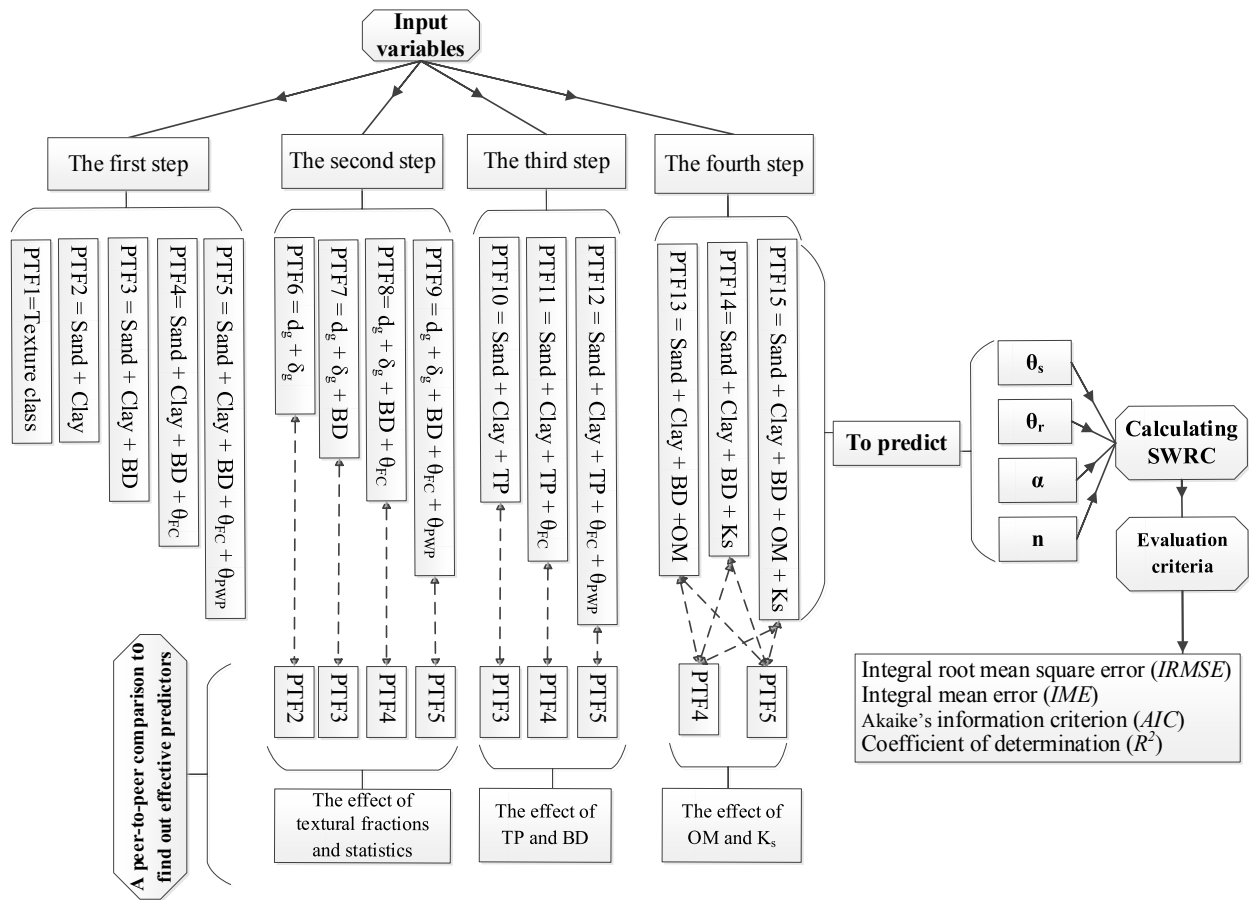
773 * Correlation is significant at the $P < 0.05$ level.

774 A list of abbreviations is available in the notation box.

775 **Fig. 5.** Results of the prediction of the soil water retention curve (SWRC) through the van
776 Genuchten model by the nonlinear regression (NLR) and random forests (RF) techniques for the
777 training step as reflected in the integral mean error (*IME*), integral root mean square error
778 (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical
779 lines indicate the standard deviations. Means with the same letter are not significantly different at
780 the significance level of $P < 0.05$ (*IRMSE* only).

781 **Fig. 6.** Results of the prediction of the soil water retention curve (SWRC) through the van
782 Genuchten model by the Rosetta software, nonlinear regression (NLR) and random forests (RF)
783 techniques for the testing step as reflected in the integral mean error (*IME*), integral root mean
784 square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion
785 (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not
786 significantly different at the significance level of $P < 0.05$ (*IRMSE* only).

787
788



789
790

791 **Fig 1.** Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten
792 model parameters (θ_r , θ_s , α and n) of the soil water retention curve (SWRC). A list of
793 abbreviations is available in the notation box.

794

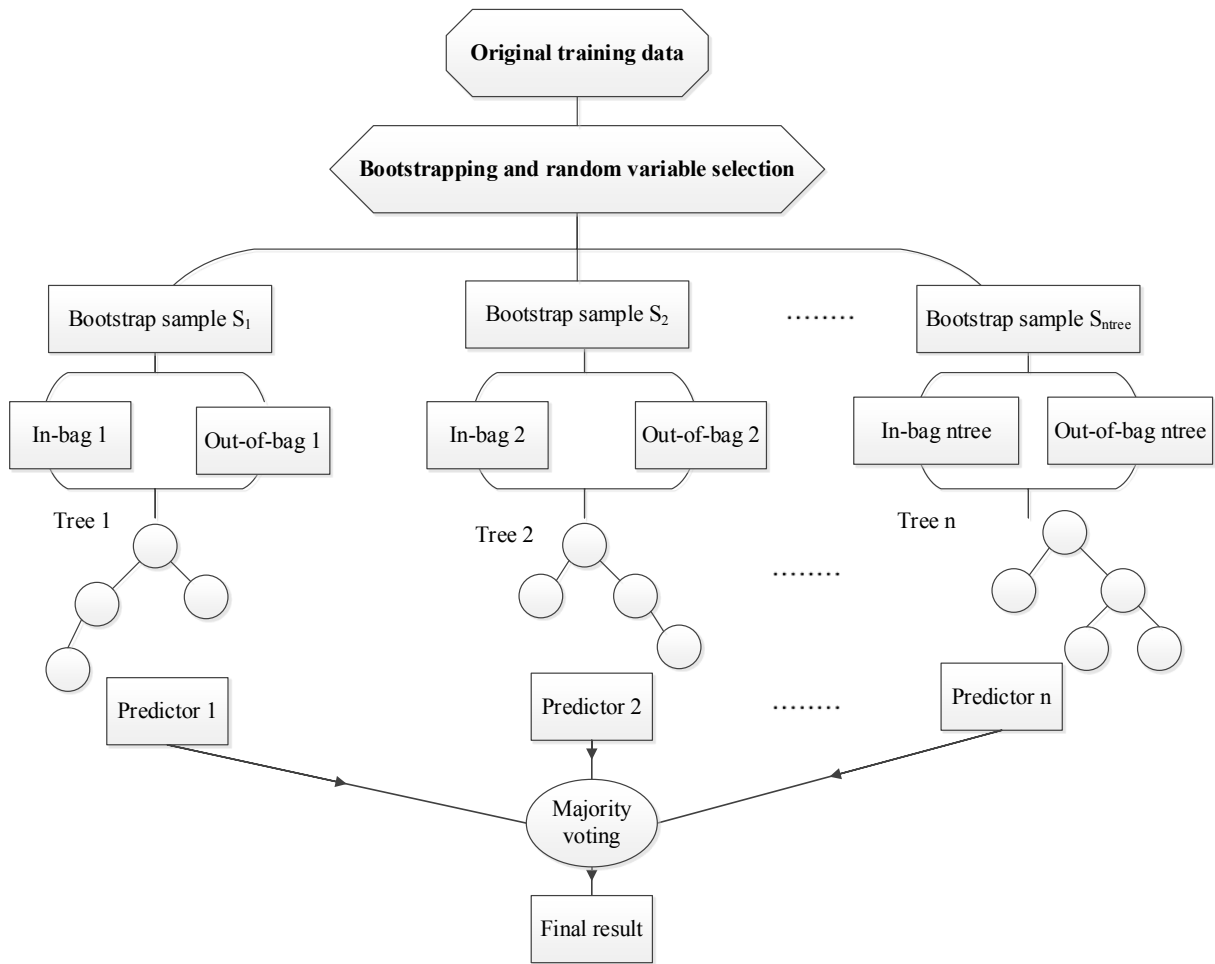
795

796

797

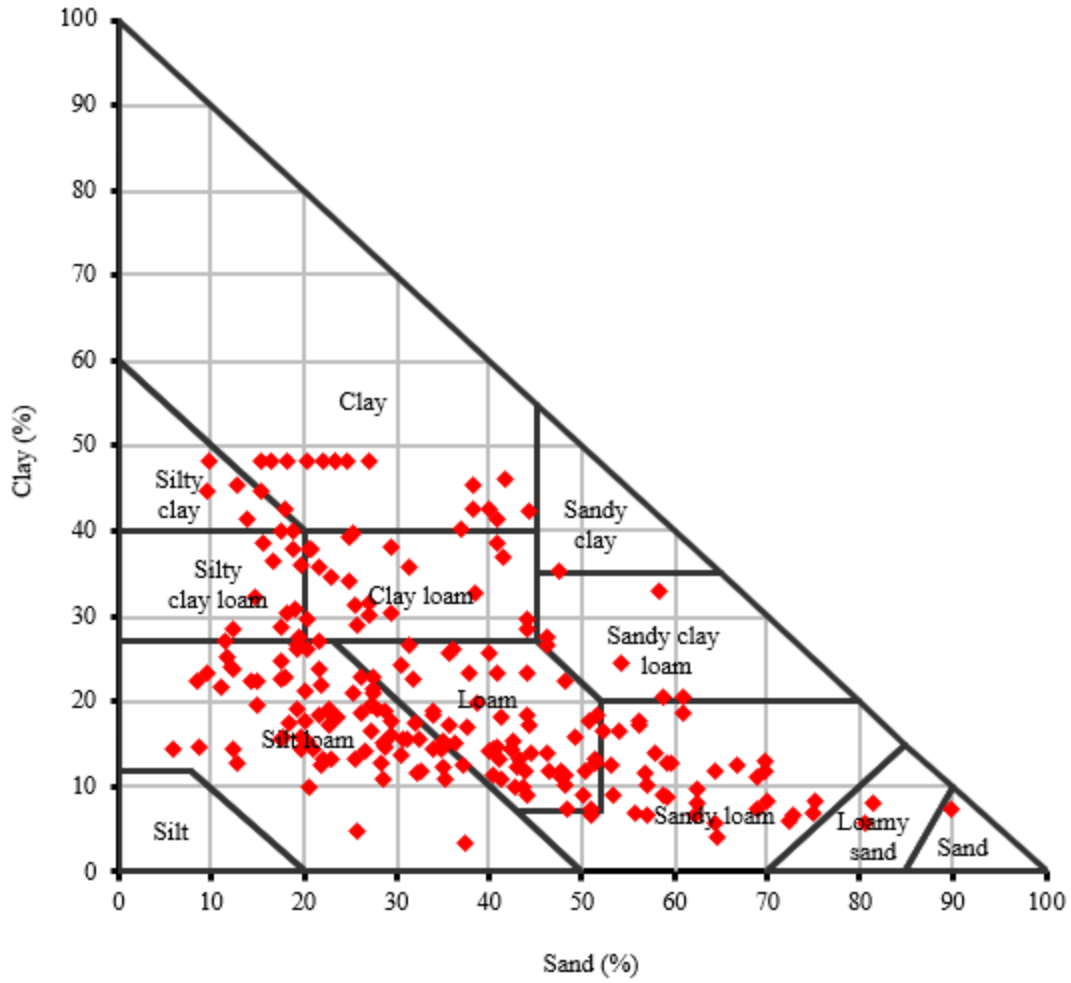
798

799



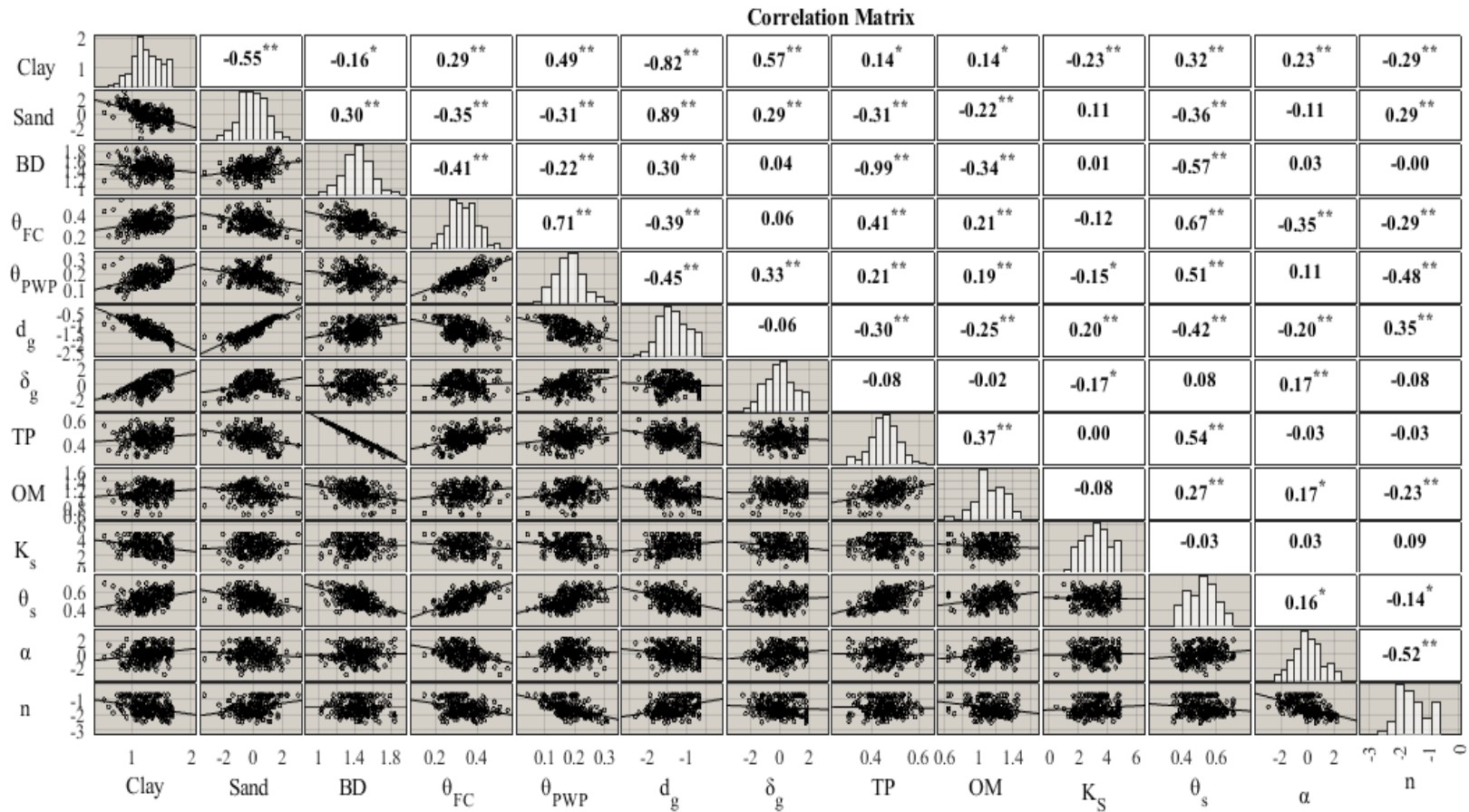
801
 802
 803
 804
 805
 806
 807
 808
 809
 810
 811
 812

Fig. 2. An architecture of a random forest.



813
 814 **Fig. 3.** Variation of soil texture classes for the dataset (n = 223) on the United States Department
 815 of Agriculture (USDA) textural triangle.

816
 817
 818
 819
 820
 821
 822

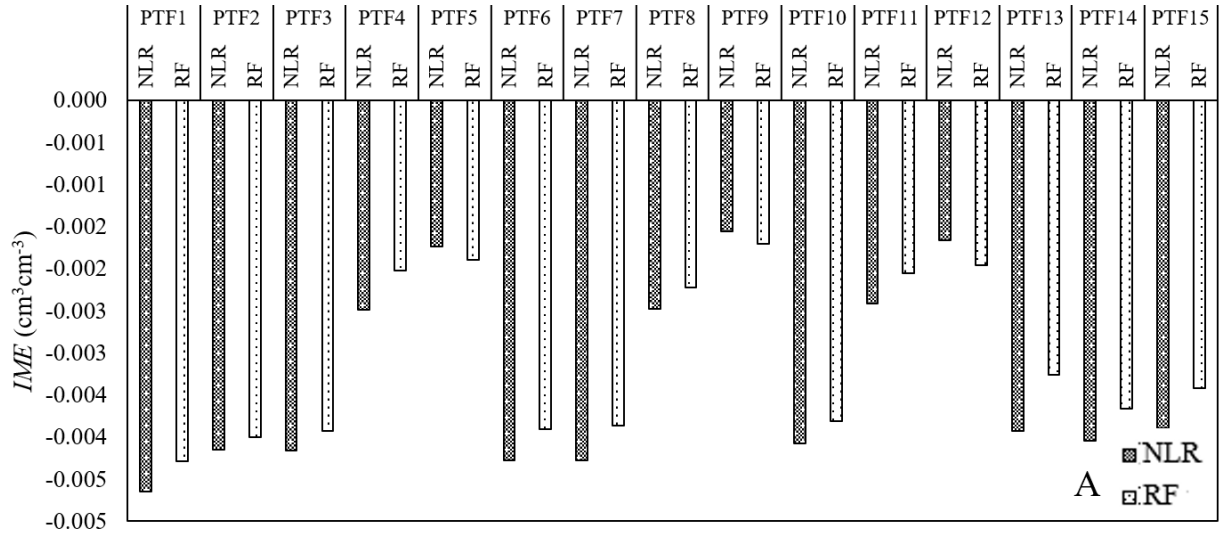


823
824 **Fig. 4.** Correlation matrix plot between input and output variables.

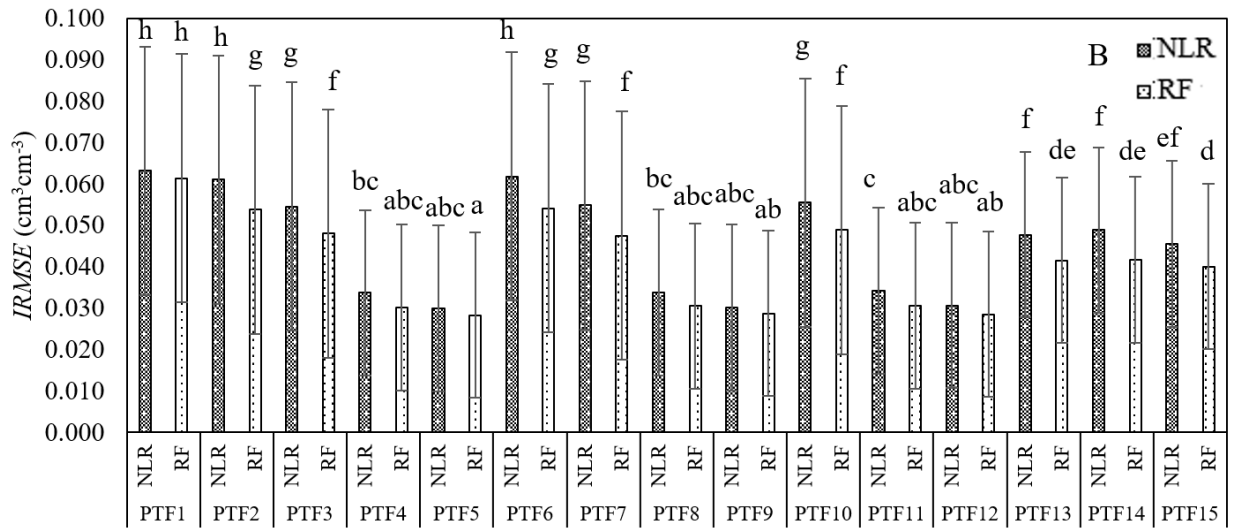
825 ** Correlation is significant at the $P < 0.01$ level.

826 * Correlation is significant at the $P < 0.05$ level.

827 A list of abbreviations is available in the notation box.



828



829

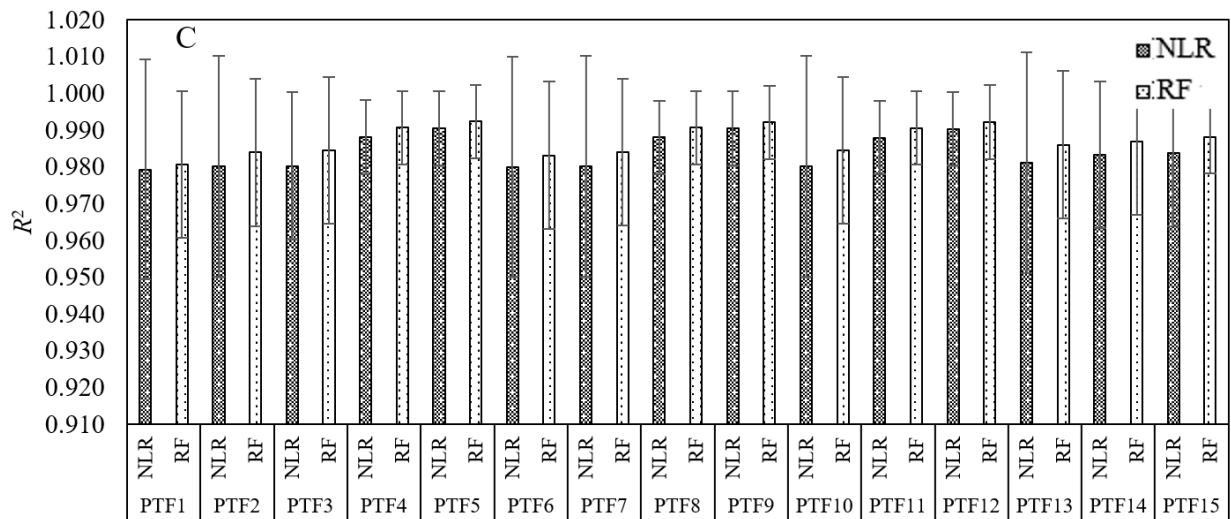
830

831

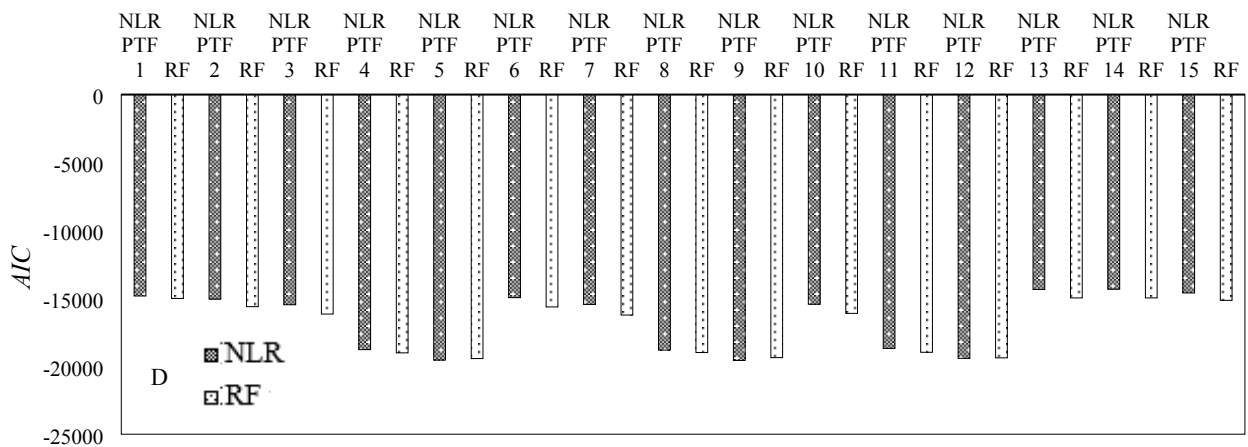
832

833

834



835



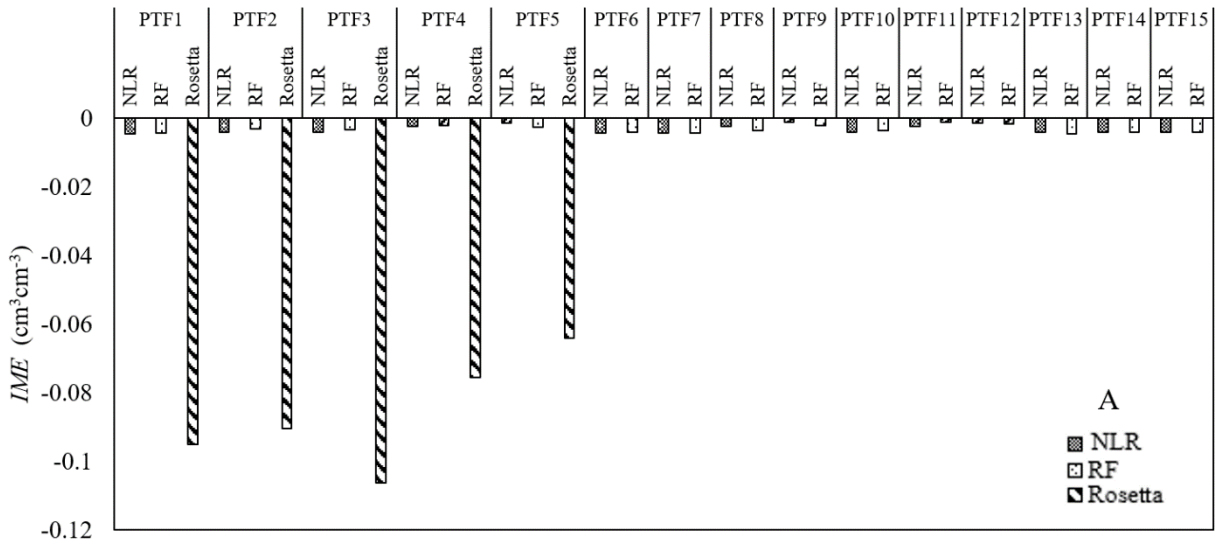
836

837

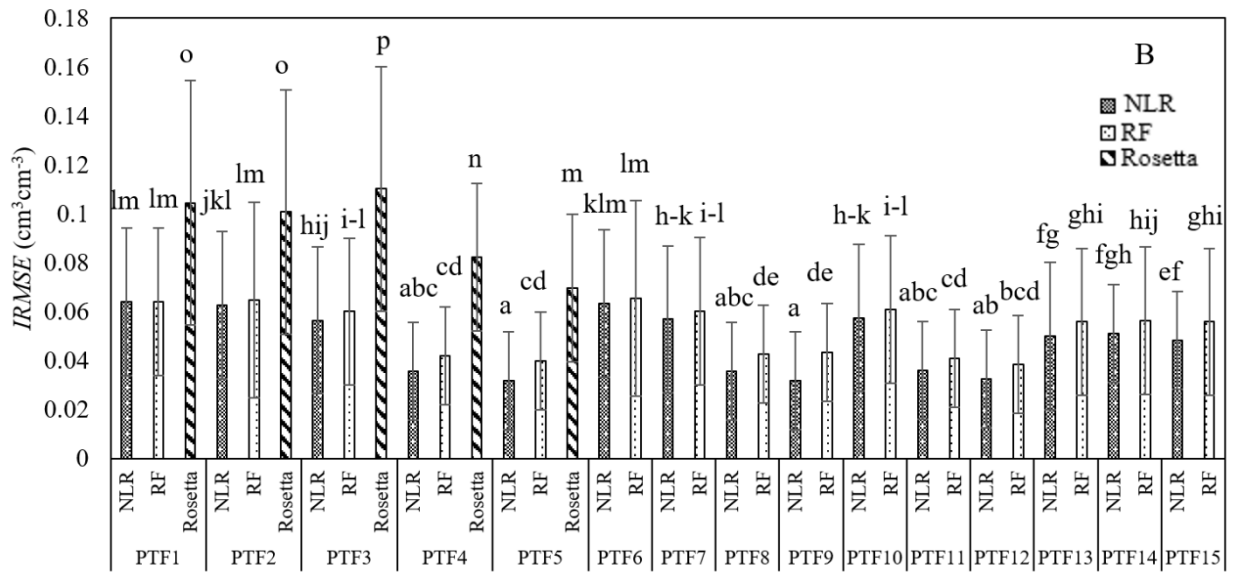
838 **Fig. 5.** Results of the prediction of the soil water retention curve (SWRC) through the van
 839 Genuchten model by the nonlinear regression (NLR) and random forests (RF) techniques for the
 840 training step as reflected in the integral mean error (*IME*), integral root mean square error
 841 (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical
 842 lines indicate the standard deviations. Means with the same letter are not significantly different at
 843 the significance level of $P < 0.05$ (*IRMSE* only).

844

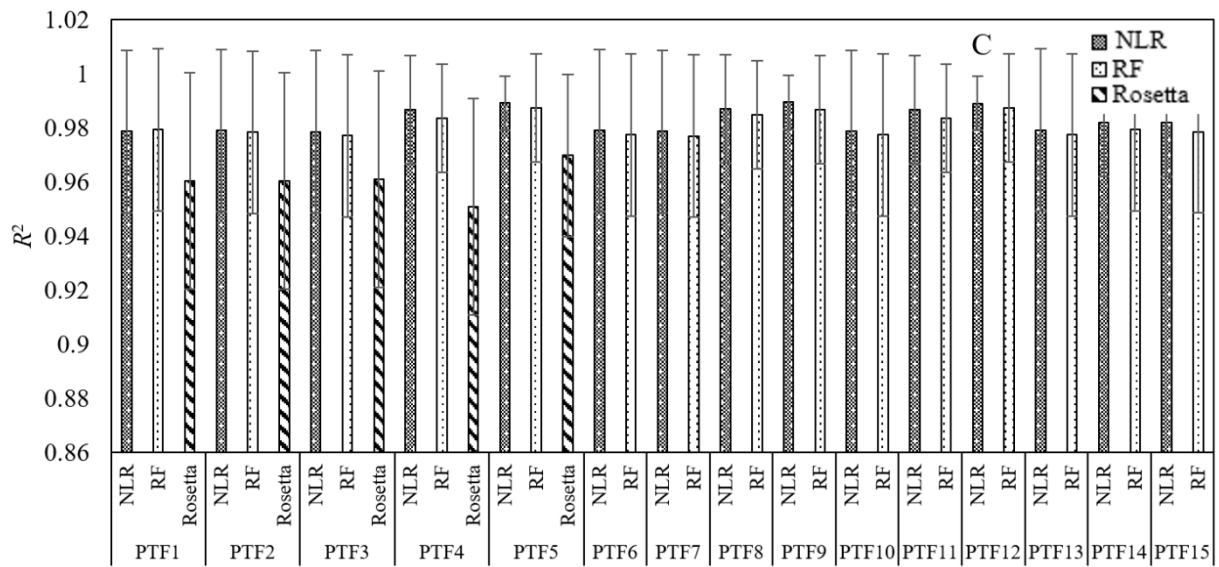
845



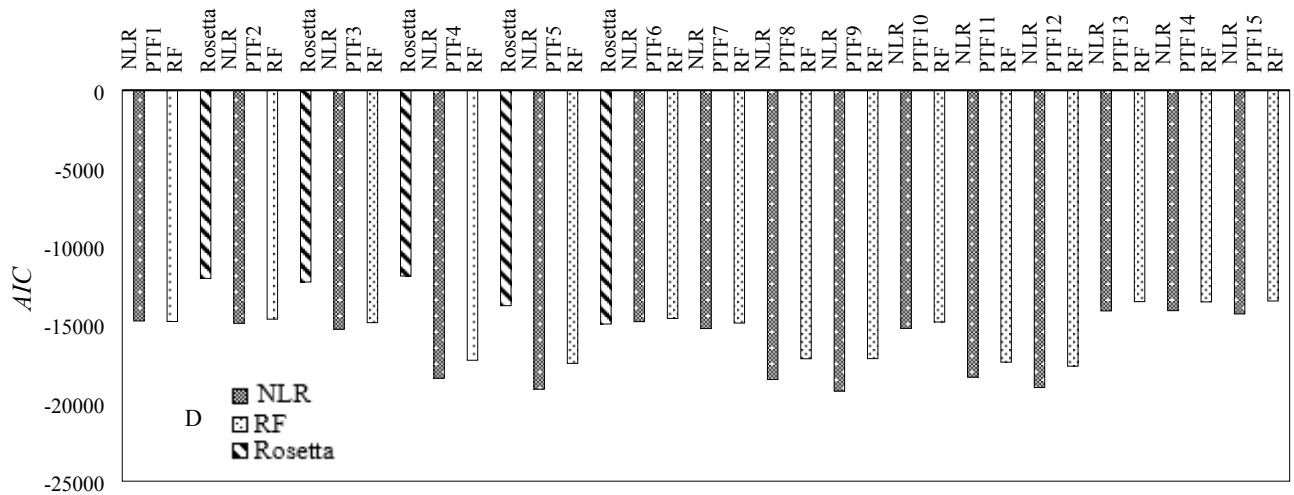
846



847



848



849

850

Fig. 6. Results of the prediction of the soil water retention curve (SWRC) through the van

851

Genuchten model by the Rosetta software, nonlinear regression (NLR) and random forests (RF)

852

techniques for the testing step as reflected in the integral mean error (*IME*), integral root mean

853

square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion

854

(*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not

855

significantly different at the significance level of $P < 0.05$ (*IRMSE* only).

856

857 **Table 1-** The results of 10, 15 and 20-fold cross-validation (k) for van Genuchten model
 858 parameters of the soil water retention curve derived from nonlinear regression (NLR) and
 859 random forest (RF) techniques based on root mean square error (*RMSE*) for pedotransfer
 860 functions PTF 3, 5 and 11 in the train and test datasets.

			θ_r			θ_s			α			n			
			<i>RMSE</i>			<i>RMSE</i>			<i>RMSE</i>			<i>RMSE</i>			
			Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	
PTF3	k=10	NLR	0.058	0.060	0.059	0.063	0.065	0.064	1.017	1.037	1.027	0.426	0.436	0.431	
		RF	0.052	0.061	0.056	0.058	0.073	0.066	0.893	1.084	0.989	0.374	0.442	0.408	
	k=15	NLR	0.058	0.060	0.059	0.064	0.064	0.064	1.017	1.030	1.024	0.426	0.434	0.430	
		RF	0.052	0.061	0.057	0.058	0.070	0.064	0.894	1.033	0.964	0.374	0.441	0.408	
	k=20	NLR	0.058	0.060	0.059	0.064	0.064	0.064	0.064	0.064	0.064	0.064	0.426	0.437	0.432
		RF	0.051	0.060	0.056	0.057	0.071	0.064	0.057	0.071	0.064	0.368	0.442	0.405	
PTF5	k=10	NLR	0.051	0.053	0.052	0.053	0.054	0.054	0.764	0.796	0.780	0.380	0.397	0.389	
		RF	0.043	0.056	0.050	0.046	0.056	0.051	0.675	0.869	0.772	0.327	0.411	0.369	
	k=15	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.764	0.790	0.777	0.381	0.399	0.390	
		RF	0.044	0.054	0.049	0.046	0.055	0.050	0.679	0.848	0.763	0.329	0.421	0.375	
	k=20	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.765	0.789	0.777	0.381	0.399	0.390	
		RF	0.042	0.054	0.048	0.044	0.054	0.049	0.654	0.842	0.748	0.316	0.412	0.364	
PTF11	k=10	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.018	1.052	1.035	0.431	0.448	0.440	
		RF	0.050	0.061	0.056	0.047	0.057	0.052	0.770	0.978	0.874	0.370	0.443	0.406	
	k=15	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.019	1.037	1.028	0.432	0.447	0.439	
		RF	0.050	0.060	0.055	0.047	0.057	0.052	0.770	1.009	0.889	0.369	0.450	0.410	
	k=20	NLR	0.058	0.060	0.059	0.065	0.065	0.065	1.020	1.024	1.022	0.432	0.439	0.435	
		RF	0.049	0.061	0.055	0.046	0.056	0.051	0.745	0.964	0.855	0.361	0.443	0.402	

861

862

863

864

865

866

867

868

869 **Table 2-** Some descriptive statistics of the measured soil variables and parameters of the van

870 Genuchten model of the soil water retention curve for the entire dataset (223 soil samples).

Variables ^a	Mean	CV (%)	Minimum	Maximum	<i>P</i> -value
Clay content (%)	21.39	54.05	3.47	48.00	0.00
Log (clay content)	1.27	19.08	0.54	1.68	0.66
Sand content (%)	35.45	48.40	5.90	89.80	0.00
Sand content*	-0.01	-14350.94	-3.40	3.14	0.90
Bulk density (g cm ⁻³)	1.43	10.97	1.03	1.84	0.83
θ_{FC} (cm ³ cm ⁻³) [§]	0.33	20.44	0.15	0.55	0.45
θ_{PWP} (cm ³ cm ⁻³)	0.18	26.21	0.04	0.31	0.90
d_g (mm)	0.07	86.62	0.00	0.21	0.00
Log (d_g)	-1.33	-27.91	-2.34	-0.67	0.77
δ_g (-)	11.57	29.39	4.54	19.97	0.00
δ_g^*	-0.01	-9872.87	-2.53	1.80	0.96
Total porosity (cm ³ cm ⁻³)	0.46	13.26	0.31	0.61	0.67
Organic matter content (%)	1.84	53.68	0.17	4.41	0.00
(Organic matter content) ^(1/4)	1.13	14.83	0.64	1.45	0.86
K_s (cm day ⁻¹)	169.10	96.58	0.06	530	0.00
$(K_s)^{(1/4)}$	3.23	30.37	0.50	4.80	0.59
θ_r (cm ³ cm ⁻³)	0.04	158.05	0.00	0.17	0.00
θ_s (cm ³ cm ⁻³)	0.52	16.26	0.35	0.70	0.56
α (kPa ⁻¹)	0.06	115.62	0.00	0.29	0.00
α^*	0.01	8889.14	-2.93	2.19	0.93
n	1.24	9.80	1.08	1.48	0.00
Ln ($n-1$)	-1.55	-30.92	-2.52	-0.74	0.05

871 ^a CV, coefficient of variation.872 [§]. A list of abbreviations is available in the notation box.873 * Normalized form of sand content: $0.91+1.06 \times \text{Ln}((\text{sand content}- 4.3)/(100.2-\text{sand content}))$;874 normalized form of δ_g : $-1.04657+1.39359 \times \text{Asinh}((\delta_g- 8.4)/3.04)$; and normalized form of α :875 $3.6+0.92 \times \text{Ln}((\alpha- 8.2 \times 10^{-6})/(1.6-\alpha))$. *P*-value is a significance value for normality test.

876

877 **Table 3-** The variance inflation factor (*VIF*) values for normalized form of the input variables.

PTFs	Clay* (%)	Sand (%)	BD ^s (g cm ⁻³)	θ_{FC} (cm ³ cm ⁻³)	θ_{PWP} (cm ³ cm ⁻³)	d_g (mm)	δ_g (-)	TP (cm ³ cm ⁻³)	OM (%)	K_s (cm day ⁻¹)
PTF2	1.42	1.42								
PTF3	1.43	1.52	1.10							
PTF4	1.45	1.56	1.25	1.31						
PTF5	1.79	1.58	1.27	2.48	2.56					
PTF6						1.00	1.00			
PTF7			1.11			1.11	1.01			
PTF8			1.25	1.33		1.01	1.22			
PTF9			1.28	2.50	2.73	1.34	1.22			
PTF10	1.55	1.43						1.11		
PTF11	1.58	1.46		1.32				1.26		
PTF12	1.60	1.79		2.49	2.56			1.28		
PTF13	1.48	1.65	1.25						1.14	
PTF14	1.55	1.64	1.14							1.06
PTF15	1.55	1.65	1.25						1.15	1.06

878 * Normalized form of the input variables is available in Table 2.

879 ^s. A list of abbreviations is available in the notation box.

880

881 **Table 4-** Analysis of variance of the integral root mean square error (*IRMSE*) of the prediction of
 882 the soil water retention curve by different methods (nonlinear regression and random forest) and
 883 pedotransfer functions (PTFs 1-15) for both the train and test datasets.

	Source	Degree freedom	Mean square	<i>F</i> -value	<i>P</i> -value
Train	Repeat (Block)	222	0.007	19.09	<0.0001
	PTFs	14	0.062	180.68	<0.0001
	Methods	1	0.038	109.69	<0.0001
	PTFs × Methods	14	0.001	1.78	0.0356
	Error	6288	0.0003		
Test	Repeat (Block)	222	0.010	16.04	<0.0001
	PTFs	14	0.073	117.22	<0.0001
	Methods	2	0.656	1056.43	<0.0001
	PTFs × Methods	18	0.002	3.68	<0.0001
	Error	7398	0.0006		

884

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Author statement:

Mostafa Rastgou:

Data curation, Writing- Original draft preparation, Visualization, Investigation, Formal analysis.

Hossein Bayat:

Conceptualization, Methodology, Writing, Supervision, Project administration, Funding acquisition.

Muharram Mansoorizadeh:

Software, Validation.

Andrew S. Gregory:

Writing- Reviewing and Editing