INTERNATIONAL Agrophysics
www.international-agrophysics.org

Int. Agrophys., 2021, 35, 73-81
doi: 10.31545/intagr/132922

Estimating parameters of empirical infiltration models from the global dataset using machine learning**

Seongyun Kim[1], *Gülay Karahan*[2], *Manan Sharma*[1], *and Yakov Pachepsky*[3]*

[1]USDA ARS EMFSL, BARC East, Bldg 201, RM 103, 10300 Baltimore Ave, Beltsville MD 20705 USA
[2]Cankiri Karatekin University, Forestry Faculty, Department of Landscape Architecture, Department of Plant Material and Cultivation, 18100 Çankırı Turkey
[3]USDA ARS EMFSL, BARC East, Bldg 177C, RM 108, 10300 Baltimore Ave, Beltsville MD 20705 USA

Received November 21, 2020; accepted January 29, 2021

A b s t r a c t. It is beneficial to develop pedotransfer relationships to estimate infiltration equation coefficients in site-specific conditions from readily available data. No systematic studies have been published concerning the relationships between the accuracy of the infiltration equation and the accuracy of the predicted coefficients in this equation. The objective of this work was to test the hypothesis that, for the same infiltration data, the accuracy of pedotransfer predictions for coefficients in an infiltration equation is greater for the infiltration equation that performs better. The hypothesis was tested using the commonly employed Horton and Mezencev (modified Kostiakov) infiltration equations with data from the Soil Water Infiltration Global database. The random forest machine learning algorithm was used to develop the pedotransfer model. The Horton and the Mezencev models performed better with 928 and 758 datasets, respectively. The accuracy of the estimates of the infiltration equation coefficients did not differ substantially between the estimates obtained from all data and from the data where the infiltration equation had lower root-mean-squared error values. The root-mean-squared error values of the pedotransfer estimates decreased by 2 to 25% when only datasets with the same infiltration measurement method were considered. The development of predictive pedotransfer equations with the data obtained from the same infiltration measurement method is recommended.

K e y w o r d s: infiltration modelling, random forest, Soil Water Infiltration Global database

*Corresponding author e-mail: Yakov.Pachepsky@usda.gov
**This work was supported by Scientific and Technological Council of Turkey (TUBITAK), grant of BIDEP (2219) (from 01.07.2019 to 07.07.2019).

INTRODUCTION

Infiltration is the key process of the hydrological cycle. Infiltration estimates are of paramount importance in flood and drought management, irrigation and drainage system design, groundwater recharge assessment, subsurface flow, and contaminant transport investigation and modelling. A large number of equations have been proposed to simulate and predict infiltration (Mishra *et al*., 2003). Both physics-based equations, *e.g.*: Brutsaert (1977), Green and Ampt (1911), Kutílek and Krejča (1987), Philip (1957), Swartzendruber (1987), and empirical equations, *e.g.* Kostiakov (1932), Horton (1940), Holtan (1961), Mezencev (1948) are in use.

Infiltration measurements are both time consuming and labour-intensive and are therefore impractical for large-scale projects. Such projects benefit from predictive models that relate the parameters of the infiltration equations to the readily available or more easily attainable site-specific data. Estimating the parameters of the infiltration equations from their soil and landscape properties has led to the development of special types of pedotransfer function (Pachepsky and Rawls, 2003). The parameters of various infiltration equations have been estimated using basic soil properties, such as clay, silt, and sand content, organic matter content, and initial soil water content (Lei *et al.*, 2020; Pandey and Pandey, 2019; Santra *et al.*, 2021; Van de Genachte *et al.*, 1996). Soil properties, which are known to be informative with regard to site-specific or region-specific conditions were often included as predictors. For example, Rahmati *et al.* (2017) included electrical conductivity and wet aggregate stability for arid soils in Iran, Brevnova (2001) added the SCS curve number for a mountainous area in the USA. Soil hydraulic parameters such as water retention parameters and hydraulic conductivity appeared to be influential predictors in the studies of Parchami-Araghi *et al.* (2013), Shao and Baumgartl (2014), and Salahou *et al.* (2020). Various vegetation-related parameters were also found to be important predictors of the parameters in infiltration equations. Shao and Baumgartl (2014) noted that each infiltration parameter was controlled by not only soil factors but also by vegetation and rainfall. It was found that soil properties alone were not sufficient to predict the infiltration parameters. Ground cover and root contents were important predictors in the works of Kidwell *et al.* (1997) and van de Genachte *et al.* (1996). Reports concerning the effect of infiltration measurement methods on the parameters of infiltration equations have been published (Maneshwari, 1997; Mazloom and Foladmand, 2013), but remain scarce.

Although the satisfactory dependencies of the parameters of infiltration equations on soil and vegetation attributes were in many cases established using linear regressions (Kidwell *et al.*, 1997; Brevnova, 2001; Shao and Baumgartl, 2014; Pandey and Pandey, 2019; Santra *et al.*, 2021), it was noted that imposing linear relationships ignores the possible nonlinearity in sought after dependencies, and may misdirect the search for the most influential predictors. Machine learning algorithms that allow for the mitigation of these problems appeared to be a suitable means for estimating the parameters of infiltration equations. Parchami-Araghi *et al.* (2013) applied artificial neural networks (ANN) to estimate the parameters of six infiltration models. Rahmati *et al.*, (2017) demonstrated the advantages of machine learning algorithms ANN and GMDH over multiple linear regression in the development of a pedotransfer relationship for parameter estimation in Kostiakov and Green-Ampt infiltration equations. Lei *et al.* (2020) applied the support vector machines (SVM) algorithm and demonstrated its advantage over ANN and linear regression.

The accuracy of the infiltration models was compared by using datasets representing local or regional conditions, it was found that the performance of the infiltration equations varied. In particular, the Horton equation performed best at 16 sites in experiments with the tillage effect concerning infiltration in Brazil (de Almeida *et al.*, 2018), in experiments involving a comparison of infiltration equations at six locations in Pakistan (Farid *et al.*, 2019), and in a 42-site study on pedotransfer function evaluation in Ethiopia (Bayabil *et al.*, 2019). The Modified Kostiakov equation known also as the Mezencev and Lostiakov-Levis equation was noted by Furman *et al.* (2006) as the most commonly used infiltration function in surface irrigation applications. The Dashtaki *et al.* (2009) comparison concluded that the Mezencev equation provided the best site-independent performance across 123 sites representing different soil series.

The pedotransfer models designed to obtain the coefficients of infiltration equations were usually developed for a single equation, and sometimes for several equations, from a single dataset obtained with a single infiltration measurement method. The performance of the infiltration equation with this dataset and the infiltration measurement method used were not considered as factors affecting the pedotransfer predictions of the infiltration equation coefficients. Our hypotheses were that: (a) the accuracy of a coefficient prediction model for a particular infiltration equation may be improved with the data with which this infiltration equation performs better, and (b) the infiltration measurement method may be an influential predictor of the infiltration equation coefficients. Our objective was to test these hypotheses using certain Horton and Mezencev infiltration equations and the large international soil infiltration database SWIG. We were also interested in analysing the input variable importance in the models for the infiltration equation parameters as determined by the random forest algorithm which was employed in this work.

MATERIALS AND METHODS

The flowchart of the modeling work is shown in Fig. 1. The data were extracted from the Soil Water Infiltration Global (SWIG) database (Rahmati *et al.*, 2018). The SWIG data were collected from 1976 to 2017. The database contains cumulative infiltration data, soil textural information, soil bulk density, organic matter content, land use, and the infiltration measurement method for 5023 datasets from 54 different countries across nearly all continents. A small number of samples have additional soil properties. Soil properties that are available from the SWIG database are summarised in Supplemental Table 1 with their statistical description. Approximately 76% of datasets contain clay, silt, and sand contents. The bulk density and organic carbon content are available in 66 and 62% of datasets, respectively. Land-use type is available in approximately 76% of datasets. In this study, 22 SWIG categories of land use types were grouped into seven categories in this work as shown in Supplemental Table 2, agriculture (cropland) is the most frequently found land use in the SWIG databases with a frequency of 53%, this is followed by grassland, pasture, garden, forest, others, and urban use.

Several methods were used to measure infiltration (Supplemental Table 3). Disc-based infiltrometers (disc, minidisc, micro-disc, Hood, and tension infiltrometers) were employed to obtain approximately 51% of the datasets. The mini disc infiltrometer is the most frequently reported infiltration method in the SWIG database with a value of about 23% (1140 out of 5023). The double ring infiltrometer is the second most frequently represented infiltration method, with 16% of datasets. The disc infiltrometer with 12% and the single ring infiltrometer with 11% are ranked as the third and fourth methods by their occurrence in the SWIG.

Two empirical equations: Horton and Mezencev were selected to evaluate their performance at simulating infiltration in this study. The infiltration model equations are listed in Table 1. Both equations are three-parametric. To avoid confusion, the parameters were renamed h_1, h_2, h_3 for Horton, and m_1, m_2, and m_3 for the Mezencev equation as shown in Table 1. Cumulative infiltration data from SWIG were used to estimate certain parameters of the Horton and Mezencev infiltration models using R version 3.53 (R core team, 2019). The NLS-search routine with mapply was used to fit the infiltration equation. Approximately 200 datasets were found in which the cumulative infiltration oscillated. Datasets with more than five oscillations were excluded before computing parameters and outliers of parameters were also removed after computing. Outliers were eliminated using the interquartile range.

The performance of the infiltration equations was evaluated using the root-mean-squared error (*RMSE*):

$$RMSE = \sqrt{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2/n},$$

where: n is the total number of observations, Y_i^{obs} is the ith observation of cumulative infiltration, and Y_i^{sim} is the ith simulation of the cumulative infiltration.

The results of fitting the Horton or Mezencev models to all datasets were referred to as H-all and M-all. H-best and M-best abbreviations were used for the results obtained from the subsets of the database for which the Horton model produced a smaller *RMSE* than the Mezencev model and vice versa, respectively. H-best and M-best were further subdivided into groups of datasets with the same measurement method. The largest number of datasets where the Horton equation performed better were obtained through the use of the minidisc infiltrometer. The largest number of datasets where the Mezencev equation performed better were obtained with the double ring infiltrometer. The abbreviation H-MDI was used for the results obtained with the minidisc infiltrometer for the Horton equation with the datasets in which the Horton equation performed better than the Mezencev equation. The abbreviation-DRI was used for the results obtained with the double ring infiltrometer for the Mezencev equation with the datasets in which the Mezencev equation performed better than the Horton equation.

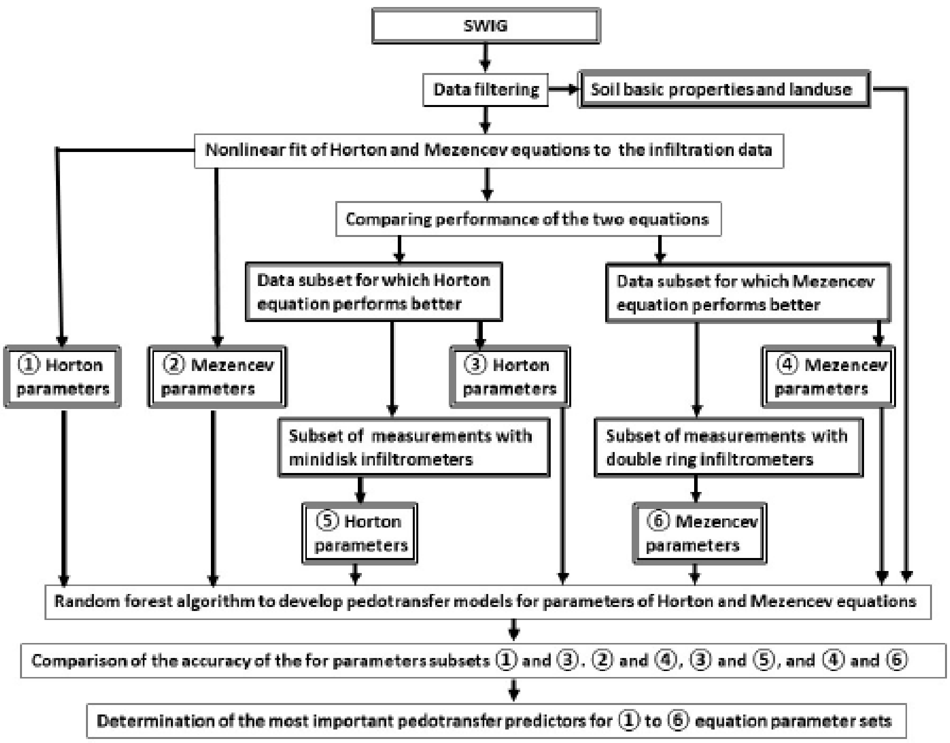

Fig. 1. Flowchart of pedotransfer modelling in this work.

In this work, the random forest algorithm (RF) was used to predict certain parameters from infiltration models. The RF is a popular machine learning algorithm for prediction and classification. It is known to be a relatively simple machine learning algorithm to train and tune (Hastie *et al.*, 2009) which builds many decision trees and averages their predictions to obtain a desirable output. In this work, RF algorithms were used as implemented in the randomForest package in R version 3.53 (Liaw and Wiener, 2018). The input variables for RF were soil textural fraction contents (clay, silt, and sand), organic carbon (OC), bulk density (Db), land use class, and infiltration measurement method. The land use and infiltration method were defined as categorical variables with 7 and 12 categories, respectively. If one of the input variables was missing in a dataset, these datasets were not used to develop the RF model. The database was split 70-30% into training and testing datasets, respectively. The default number of trees (500) was applied.

The input variable importance was measured using the mean decrease accuracy (%IncMSE) as implemented in the R randomForest package. The Mean Decrease Accuracy (%IncMSE) reflects the loss in model accuracy when the variable is scrambled, *i.e.* its values are randomly replaced with values that have the same statistical distribution. The model decrease in accuracy computed for each tree in the forest and the percentage decrease in accuracy is averaged over all trees in order to obtain the mean value.

RESULTS

The cumulative distribution functions (CDFs) of fitted parameters are shown in Fig. 2. The CDF of the Horton model parameter h_1 have similar patterns for H-all, H-best, and H-MDI datasets. The median value ranged from 0.52 for H-best to 0.70 for H-MDI and the standard deviations ranged from 0.50 for H-MDI to 0.64 for H-best. Whereas the CDF of parameters h_2 and h_3 show similar patterns for H-all and H-best datasets, h_2 and h_3 CDFs for the H-MDI dataset have shapes that are different from those for H-all and H-best, and there is less variability in h_2 and h_3 for the H-MDI datasets. The standard deviations in logarithm value were 0.63 of the -MDI in parameter h_3 in CDFs for the H-all, H-best, and H-MDI datasets, respectively. While the CDFs of each parameter of the Mezencev equation were similar across the subsets of M-all, M-best and M-DRI, the median value in parameter m_1 for the M-DRI dataset was slightly less than parameter m_1 at M-all and M-best. Only 2% of the fitted values of m_2 were larger than 1.0, which indicated the concave shapes of the cumulative infiltration curves. The other 98% of the datasets were convex with m_2>1 as envisaged in the Mezencev (1948) work.

The root-mean-squared errors of the random forest models developed for parameter estimation are given in Table 2. The performance of the parameter estimation models in terms of *RMSE* values improved only slightly as the estimation was carried out only for datasets where the infiltration equations were performing better than their counterparts. The *RMSE* values of h_1, h_2, and h_3 estimates for the H-best datasets were lower than those of the H-all

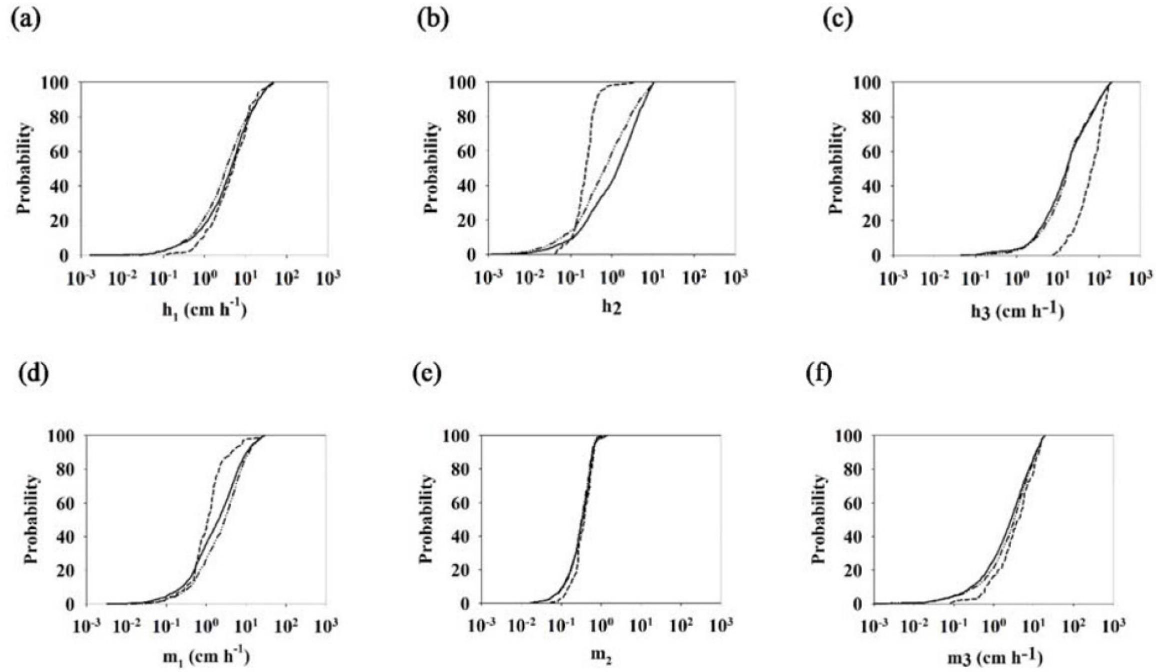

Fig. 2. Cumulative distribution functions of the fitting parameters from the Horton and Mezencev infiltration equation; ——— H-all or M-all, —··—··— H-best or M-best, — — — H-method or M-method: (a) parameter h_1 from the Horton model, (b) parameter h_2 from the Horton model, (c) parameter h_3 from the Horton model, (d) parameter m_1 from the Mezencev model, (e) parameter m_2 from the Mezencev model and (f) parameter m_3 from the Mezencev model.

Table 1. Infiltration equations used in this study

Equation	Infiltration equations	
	Original form of the equation	Equation forms in this study
Horton	$F(t) = f_c t + \frac{f_0 - f_c}{k}\left(1 - e^{-kt}\right)$	$F(t) = h_3 t + h_1\left(1 - e^{-h_2 t}\right)$
Mezencev	$F(t) = kt^a + f_0 t$	$F(t) = m_1 t^{m_2} + m_3 t$

$F(t)$ – cumulative infiltration (cm) at time t (h). In the Horton Equation, f_c – final or equilibrium infiltration rate (cm h^{-1}), f_0 – initial infiltration rate (cm h^{-1}), k – constant representing the exponential rate of decrease of infiltration (h^{-1}). In the Mezencev equation, k (cm h^{-a}), a, unitless and f_0 (cm h^{-1}) are empirical constants ($k > 0$ and $0 < a < 1$) for the Mezencev equation, h_1, h_2, h_3 and m_1, m_2, m_3 are fitting parameters in this study.

datasets. Similarly, the *RMSE* values of m_1, m_2, and m_3 estimates for the M-best dataset were lower than those of the M-all dataset. A substantial decrease in *RMSE* occurred when the only datasets that were used were the ones for which (a) the equation performed better, and (b) the infiltration measurement method was the same.

In this case, the *RMSE* values of h_1, h_2 and h_3 decreased by 15, 22, and 6% and the *RMSE* values of m_1, m_2, and m_3 decreased by 2, 14, and 25%, respectively.

The one-to-one scatterplot comparison between the fitted and the estimated with random forest parameters of infiltration equations is shown in Fig. 3. These data are also characterised in the Supplemental Table S4 containing the R^2 values. When H-best is considered rather than H-all, R^2 of the parameter h_1 estimation result increases and R^2 of the h_2 and h_3 parameters decreases. Similarly, when M-best is considered instead of M-all, R^2 of the parameter m_2 estimation result increases and R^2 of the m_1 and m_3 parameters decreases. In the majority of cases, the R^2 values of the parameter estimates with datasets for specific methods are low because the range of parameter variation is comparable with the range of the estimation error variation (Fig. 2). The R^2 value does not characterise the differences in the accuracy of the estimates in this case.

The relative predictor importance ranked in terms of the Mean Decrease in Accuracy is shown in Table 3. Only the top three important predictors are listed. Infiltration measurement methods were the most important predictors for all of the parameters of both the Horton and Mezencev equations. Infiltration measurement were first ranked in terms of estimating all of the parameters from H-all, M-all, H-best, and M-best datasets. The second most important predictors were the soil textural fractions (clay, sand, silt). The clay content achieved a slightly higher rank as a more important variable than sand and silt in all parameters of the Horton equation. In the estimation of h_3, the bulk density was ranked in second place in the estimation scheme of H-all and third in the estimation scheme of H-best. Soil texture was found to be the most important predictor of the m_3 parameter in the Mezencev equation. In the case of estimations for a specific measurement method with H-MDI and M-DRI datasets, in which the infiltration method was not included as the predictor, the organic carbon content became one of the important predictors.

DISCUSSION

A comparison of the *RMSE* values of the parameter predictive models showed that homogeneous datasets in terms of the model performance did not provide more accurate estimations, however, performance was improved for datasets that were homogeneous in terms of the measurement method (Table 2). There may be several reasons for the influence of the measurement method. Soil surface preparation could be one of them. For example, Shao and Baumgartl (2014) compared ring infiltrometer and sprinkler infiltrometer measurements and noted that both the vegetation and surface sealing effects from raindrops were both affecting the infiltration measurements in rainfall simulation and were neglected in ring infiltrometry since the latter is commonly applied on soil stripped of vegetation and a levelled ground surface.

Another reason for the measurement method being among the most important predictors of the scale effect arises from the difference in the areas of contact surfaces between the infiltration measurement methods. For example, the contact areas are 16 and 700 cm^2 for the minidisc infiltrometer and double-ring infiltrometer, respectively. The infiltration flow occurs in different volumes and different horizons of soils, and the flow from different contact areas encounters different levels of soil structural heterogeneity. Previous studies showed that the contact area greatly affected the hydraulic conductivity measurements (Pachepsky *et al*. 2014); as the flow domain cross-section increased, the hydraulic conductivity could increase by one or two orders of magnitude and then stabilise. The pedotransfer functions for hydraulic conductivity improved when the contact area was included in the predictor list (Ghanbarian *et al* 2015). It appears that the contact area greatly influences not only the stationary stage of infiltration (from which the hydraulic conductivity value is derived) but also the parameters of the non-stationary phase.

The dimensionality of the flow domain in the soil may be yet another reason for the influence of the infiltration measurement method on the predictions of the parameters

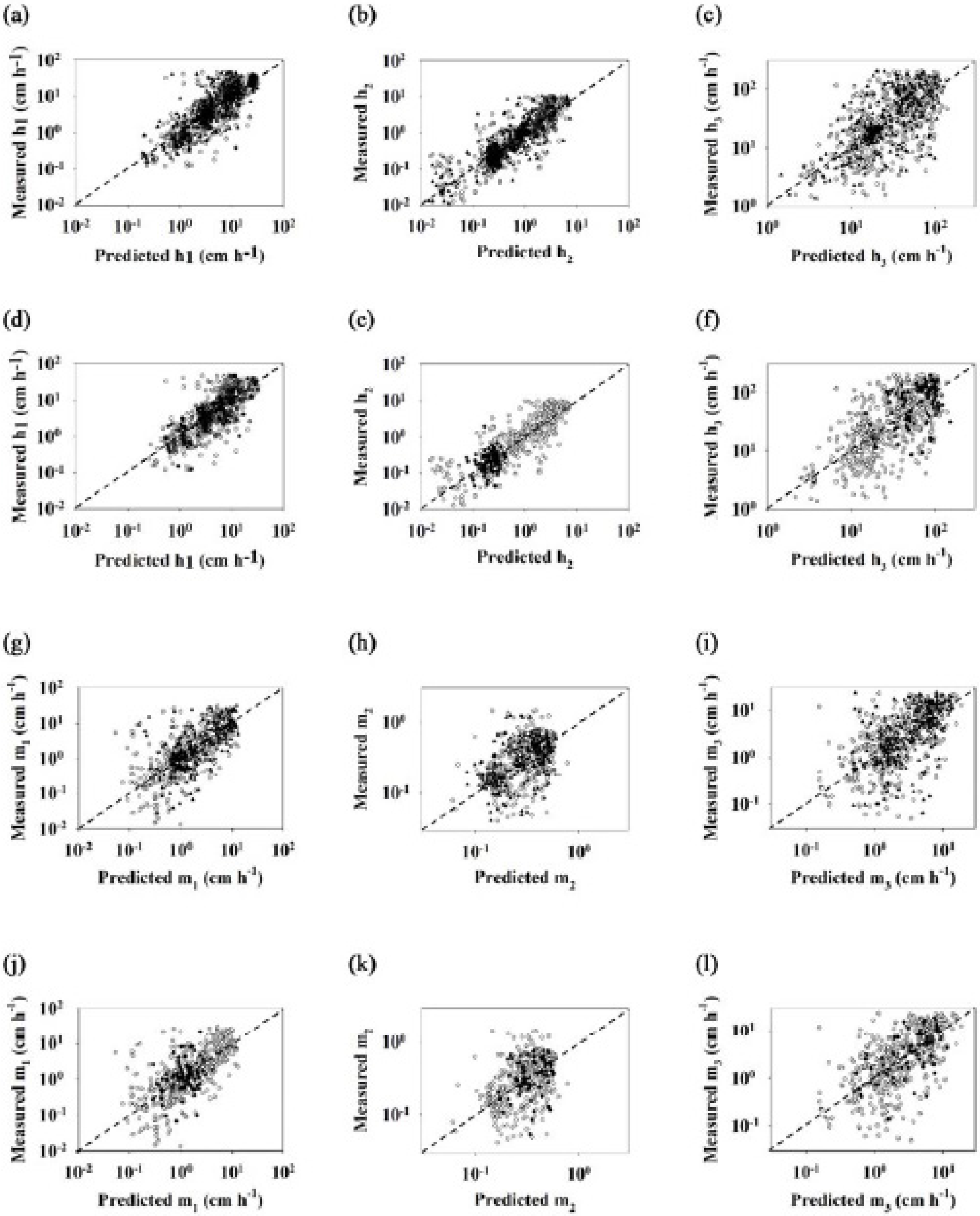

Fig. 3. Comparison of fitted and estimated with random forest model parameters of infiltration equations; a, b, c – Horton equation for all database (H-all, ○) and for the datasets where Horton equation performed better than the Mezencev equation (H-best, ▲); d, e, f – Horton equation for all database (H-all, ○) and for the subset of H-best with mini disc infiltrometer measurements only (H-MDI, ●); g, h, i – Mezencev equation for all database (M-all, ○) and for the datasets where Mezencev equation performed better than the Mezencev euqation (M-best, ▲); j, k, l – Mezencev equation for all database (M-all, ○) and for the subset of M-best with double ring infiltrometer measurements only (M-DRI, ●).

of the infiltration equations. Rahmati *et al*. (2018) noted that the double-ring infiltrometer data in SWIG could be considered one-dimensional whereas many other methods provided 3D data. These authors suggested using different infiltration equations for different dimensionality of flow in the infiltration measurements. The practical aspect of the influential effect of the infiltration method on the predictions of parameters of the infiltration equations appears to be the need to develop different measurement-method specific predictive models for the infiltration equation coefficients.

When the same method was used, the soil textural fractions and organic carbon content became the most important predictors (Table 3). It is interesting to note that

Table 2. Root-mean-squared errors of logarithms of parameters for Horton and Mezencev equations estimated using random forest modeling

Dataset	Horton equation $F(t) = h_3t + h_1 (1 - e^{-h_2 t})$		
	h_1	h_2	h_3
H-all (958)*	0.388	0.361	0.330
H-best (504)	0.372	0.347	0.328
H-MDI (142)	0.317	0.270	0.306
	Mezencev equation $F(t) = m_1 t^{m_2} + m_3t$		
	m_1	m_2	m_3
M-all (728)	0.456	0.275	0.487
M-best (378)	0.451	0.275	0.474
M-DRI (45)	0.441	0.238	0.355

*Total number of measurements in the dataset.

the soil bulk density was not in the list of the most influential inputs. It is possible that the relatively small sample taken to measure bulk density does not reflect the level of heterogeneity encountered by water flow in the double ring in the infiltrometer, and that it does not reflect the possibilities for the distribution of water between vertical and lateral flows in the measurements with the minidisc infiltrometer. The presence of organic carbon in the list of the most important predictors is expected since this is the available input that is most closely related to soil structure. Organic carbon is one of the most important predictors in the models for the parameters h_1, h_2, and m_2 which are responsible for the initial part of the infiltration curve.

The absence of land use type in the list of the most important predictors was not expected since its importance has been emphasised in several previous studies (Van de Genachte *et al.*, 1996; Kidwell *et al.*, 1997; Shao and Baumgartl, 2014). However, these studies were performed on a relatively small scale. The global dataset in SWIG may allow for such a wide variety of soil conditions for the same land use category that the value of land use alone as a predictor becomes less significant. The infiltration rate should be affected by the initial water content in the soil, the water content of which was an important variable for predicting hydraulic conductivity (Araya and Ghezzehei, 2019). Since the initial soil water content is only available in 31% of the infiltration data in the SWIG database, initial soil water content was not included in this study.

The procedure for the comparison of model performance could be one of the reasons for the lack of substantial model performance improvement after the selection of the data subset with which one model performed better than the others. A simple comparison of *RMSE* values does not reveal whether or not the difference in performance is statistically significant. Information concerning the uncertainty in the data is required to establish thresholds for the differences in *RMSE* above which the performance of the models would be significantly different. The values of *RMSE* for the parameters h_3 and m_3 which are responsible for the stationary portion of the cumulative infiltration curve, are lower than the *RMSE* estimates of the hydraulic conductivity of the soil (Pachepsky and Park, 2015; Arays and Ghezzehei, 2019). In general, the accuracy of the parameter estimation models (Table 2) cannot be evaluated without reference to

Table 3. Relative importance of the top three predictors from each parameter of Horton and Mezencev infiltration models based on Mean Decrease Accuracy H-method measured by the mini-disk infiltrometer and M-method measured by the double-ring infiltrometer

Dataset	Horton equation $F(t) = h_3t + h_1 (1 - e^{-h_2 t})$								
	h_1			h_2			h_3		
H-all	Method	Clay	OC	Method	Sand	Clay	Method	Db	Silt
	46	35	31	90	32	30	69	34	29
H-best	Method	Clay	Sand	Method	Clay	Sand	Method	Silt	Db
	40	33	31	57	27	23	50	23	22
H-MDI	Clay	OC	Silt	Silt	Sand	OC	Sand	Clay	OC
	30	23	15	18	11	10	18	15	11
	Mezencev equation $F(t) = m_1 t^{m_2} + m_3t$								
	m_1			m_2			m_3		
M-all	Method	Sand	Clay	Method	Silt	OC	Method	Clay	Sand
	65	24	22	45	22	19	36	27	26
M-best	Method	Sand	Silt	Method	Db	Clay	Method	Clay	Sand
	57	16	15	26	16	15	36	23	22
M-DRI	Sand	Silt	Clay	OC	Land	Silt	Silt	Sand	Clay
	12	11	9	15	12	8	21	12	11

their future applications. The values of *RMSE* will serve to quantify the degree of uncertainty and can be used in hydrological computations to establish the uncertainty of the simulated target values of storage, flux and carrying capacity of the water.

CONCLUSIONS

We estimated the parameters of the Horton and Mezencev infiltration equations as they are affected by soil properties, land use category, and the infiltration measurement method of 1850 datasets from the Soil Water Infiltration Global database. The application of the random forest algorithm led to the following conclusions:

1. The infiltration measurement method was by far the most important predictor of parameters followed by soil texture and organic carbon.

2. The accuracy of the predictions was moderate.

3. The accuracy of parameter estimation in the infiltration equations did not reflect the accuracy of the infiltration data approximation with these equations.

4. The functional evaluation of the predictive models should be performed before using them in the relevant application.

5. The creation of the predictive equations for specific infiltration methods may improve the accuracy of the infiltration parameter estimation.

Conflict of interest: The authors declare no conflict of interest.

REFERENCES

Araya S.N. and Ghezzehei T.A., 2019. Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. Water Res. Res., 55(7), 5715-5737. https://doi.org/10.1029/2018wr024357

Bayabil H.K., Dile Y.T., Tebebu T.Y., Engda T.A., and Steenhuis T.S., 2019. Evaluating infiltration models and pedotransfer functions: implications for hydrologic modeling. Geoderma, 338, 159-169. https://doi.org/10.1016/j.geoderma.2018.11.028

Brevnova E.V., 2001. Green-Ampt infiltration model parameter determination using SCS curve number (CN) and soil texture class, and application to the SCS runoff model. Graduate Theses, Dissertations, and Problem Reports, 1152. https://researchrepository.wvu.edu/etd/1152

Brutsaert W., 1977. Vertical infiltration in dry soil. Water Res. Res., 13(2), 363-368. https://doi.org/10.1029/wr013i002p00363

Dashtaki S.G., Homaee M., Mahdian M.H., and Kouchakzadeh M., 2009. Site-dependence performance of infiltration models. Water Res. Manag., 23(13), 2777-2790. https://doi.org/10.1007/s11269-009-9408-3

de Almeida W.S., Panachuki E., de Oliveira P.T.S., da Silva Menezes R., Sobrinho T.A., and de Carvalho D.F., 2018. Effect of soil tillage and vegetal cover on soil water infiltration. Soil Till. Res., 175, 130-138. https://doi.org/10.1016/j.still.2017.07.009

Farid H.U., Mahmood-Khan Z., Ahmad I., Shakoor A., Anjum M.N., Iqbal M.M., Mubeen M., and Asghar M., 2019. Estimation of infiltration models parameters and their comparison to simulate the onsite soil infiltration characteristics. Int. J. Agric. Biol. Eng., 12(3), 84-91. https://doi.org/10.25165/j.ijabe.20191203.4015

Furman A., Warrick A.W., Zerihun D., and Sanchez C.A., 2006. Modified Kostiakov infiltration function: Accounting for initial and boundary conditions. J. Irrig. Drain. Eng., 132(6), 587-596. https://doi.org/10.1061/(asce)0733-9437(2006)132:6(587)

Ghanbarian B., Taslimitehrani V., Dong G., and Pachepsky Y.A., 2015. Sample dimensions effect on prediction of soil water retention curve and saturated hydraulic conductivity. J. Hydrol., 528, 127-137. https://doi.org/10.1016/j.jhydrol.2015.06.024

Green W.H. and Ampt G.A., 1911. Studies on soil physics. Part I. The flow of air and water through soils. J. Agric. Sci., 4, 1-24.

Hastie T., Tibshirani R., and Friedman J., 2009. The elements of statistical learning: data mining, inference and prediction. Springer Science+Business Media, New York, NY. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Holtan H.N., 1961. A Concept for Infiltration Estimates in Watershed Engineering. USDA Bulletin, Washington, DC, USA.

Horton R.E., 1940. An approach towards a physical interpretation of infiltration capacity. Soil Sci. Soc. Am. J., 5, 399-417.

Kidwell M.R., Weltz M.A., and Guertin D.P., 1997. Estimation of green-Ampt effective hydraulic conductivity for rangelands. Rangeland Ecol. Manag./J. Range Manag. Archives, 50(3), 290-299. https://doi.org/10.2307/4003732

Kostiakov A.N., 1932. On the dynamics of the coefficients of water percolation in soils and on the necessity of studying it from a dynamic point of view for purpose of amelioration. Trans. Sixth Comm. Int. Soc. Soil Sci., 1, 7-21.

Kutílek M. and Krejča M., 1987. A three-parameter infiltration equation of the Philip's type solution (in Czech). Vodohosp. Čas., 35, 52-61.

Lei G., Fan G., Zeng W., and Huang J., 2020. Estimating parameters for the Kostiakov-Lewis infiltration model from soil physical properties. J. Soils Sediments, 20(1), 166-180. https://doi.org/10.1007/s11368-019-02332-4

Liaw A. and Wiener M., 2018. Breiman and Cutler's Random Forests for Classification and Regression. R Package 'random Forest. https://cran.r-Fproject.org/web/packages/randomForest/index.html

Maheshwari B.L., 1997. Interrelations among physical and hydraulic parameters of non-cracking soils. J. Agric. Eng. Res., United Kingdon, 68(4), 297-309.

Mazloom H. and Foladmand H., 2013. Evaluation and determination of the coefficients of infiltration models in Marvdasht region, Fars province. Int. J. Advanced Biolog. Biomedical Res., 1(8): 822-829.

Mezencev V.J., 1948. Theory of formation of the surface runoff (in Russian). Meteorol. Gidrol., 3, 33-40.

Mishra S.K., Tyagi J.V., and Singh V.P., 2003. Comparison of infiltration models. Hydrol. Process., 17(13), 2629-2652. https://doi.org/10.1002/hyp.1257

Pachepsky Y., Guber A.K., Yakirevich A.M., McKee L., Cady R.E., and Nicholson T.J., 2014. Scaling and pedotransfer in numerical simulations of flow and transport in soils. Vadose Zone J., 13(12). https://doi.org/10.2136/vzj2014.02.0020

Pachepsky Y. and Park Y., 2015. Saturated hydraulic conductivity of US soils grouped according to textural class and bulk density. Soil Sci. Soc. Am. J., 79(4), 1094-1100. https://doi.org/10.2136/sssaj2015.02.0067

Pachepsky Y. and Rawls W.J., 2003. Soil structure and pedotransfer functions. Eur. J. Soil Sci., 54(3), 443-452. https://doi.org/10.1046/j.1365-2389.2003.00485.x

Pandey P.K. and Pandey V., 2019. Estimation of infiltration rate from readily available soil properties (RASPs) in fallow cultivated land. Sust. Water Res. Manag., 5(2), 921-934. https://doi.org/10.1007/s40899-018-0268-y

Parchami-Araghi F., Mirlatifi S.M., Dashtaki S.G., and Mahdian M.H., 2013. Point estimation of soil water infiltration process using Artificial Neural Networks for some calcareous soils. J. Hydrol., 481, 35-47. https://doi.org/10.1016/j.jhydrol.2012.12.007

Philip J.R., 1957. The theory of infiltration: The infiltration equation and its solution. Soil Sci., 83, 345-357. https://doi.org/10.1097/00010694-195705000-00002

R Core Team. R, **2019.** The R project for statistical computing. http://www.R-project.org/

Rahmati M., 2017. Reliable and accurate point-based prediction of cumulative infiltration using soil readily available characteristics: a comparison between GMDH, ANN, and MLR. J. Hydrol., 551, 81-91. https://doi.org/10.1016/j.jhydrol.2017.05.046

Rahmati M., Weihermüller L., Vanderborght J., Pachepsky Y.A., Mao L., Sadeghi S.H., Moosavi N., Kheirfam H., Montzka C., Van Looy K., and Toth B., 2018. Development and analysis of the Soil Water Infiltration Global database. Earth System Science Data, 10, 1237-1263.

Salahou M.K., Jiao X., and Lü H., 2020. Assessment of empirical and semi-empirical models for estimating a soil infiltration function. Trans. ASABE, 63(4), 833-845. https://doi.org/10.13031/trans.13639

Santra P., Kumar M., and Kumawat R.N., 2021. Characterization and modeling of infiltration characteristics of soils under major land use systems in Hot Arid Region of India. Agric. Res., 1-17. https://doi.org/10.1007/s40003-020-00511-1

Shao Q. and Baumgartl T., 2014. Estimating input parameters for four infiltration models from basic soil, vegetation, and rainfall properties. Soil Sci. Soci. Am. J., 78(5), 1507-1521. https://doi.org/10.2136/sssaj2014.04.0122

Swartzendruber D., 1987. A quasi-solution of Richards' Equation for the downward infiltration of water into soil. Water Res. Res., 23(5), 809-817. https://doi.org/10.1029/wr023i005p00809

Van de Genachte G., Mallants D., Ramos J., Deckers J.A., and Feyen J., 1996. Estimating infiltration parameters from basic soil properties. Hydrol. Processes, 10(5), 687-701. https://doi.org/10.1002/(sici)1099-1085(199605)10:5<687::aid-hyp311>3.0.co;2-p

SUPPLEMENTUM

Table S1. Soil properties, number of data entries in the SWIG database (out of 5023 in total), and their statistical description (Rahmati *et al*., 2018)

Soil properties	Availability	Fr (%)	Mean	Min	Max	Median	CV (%)
Clay (%)	3842	76	24	0	80	20	64
Silt (%)	3842	76	36	0	82	37	52
Sand (%)	3842	76	41	1	100	38	63
Bulk density (g cm^{-3})	3295	66	1.32	0.14	2.81	1.35	20
Organic carbon (%)	3102	62	3	0	88	1	200

Fr - frequency (%), Min - minimum, Max - maximum, CV - coefficient of variation.

Table S2. Land use type of soils (modified from Rahmati *et al*., 2018)

Land use	Frequency	Land use	Frequency
Agriculture	2019	Forest	204
Grass	933	Others	122
Pasture	233	Urban	103
Garden	216		

Table S3. Infiltration methods used to measure infiltration (from Rahmati *et al*., 2018)

Method	Number of datasets	Method	Number of datasets
Double ring infiltrometer	828	Guelph permeameter	181
Single ring infiltrometer	570	Aardvark permeameter	50
Disc infiltrometer	607	Rainfall simulator	374
Mini disc infiltrometer	1140	Linear source method	10
Micro infiltrometer	36	Point source method	4
Hood infiltrometer	23	Beerkan(Best)	197
Tension infiltrometer	752	Not reported	251

Table S4. R^2 values of parameters for the Horton and Mezencev equations estimated using random forest modelling

Dataset	Horton equation	$F(t) = h_3 t + h_1 \left(1 - e^{-h_2 t}\right)$	
	h_1	h_2	h_3
H-all (958)*	0.569	0.757	0.532
H-best (504)	0.586	0.746	0.451
H-MDI (142)	0.601	0.004	0.358
	Mezencev equation	$F(t) = m_1 t^{m_2} + m_3 t$	
	m_1	m_2	m_3
M-all (728)	0.487	0.203	0.377
M-best (378)	0.509	0.282	0.388
M-DRI (45)	0.542	0.220	0.513

*Total number of measurements in the dataset.