The site-specific selection of the infiltration model based on the global dataset and random forest algorithm

Seongyun Kim\textsuperscript{1,2} | Gülay Karahan\textsuperscript{3} | Manan Sharma\textsuperscript{1} | Yakov Pachepsky\textsuperscript{1}

\textsuperscript{1} USDA-ARS, Environmental Microbial and Food Safety Lab., Beltsville, MD 20705, USA
\textsuperscript{2} Oak Ridge Institute of Science and Engineering, ARS Research Participation Program, Oak Ridge, TN 37831, USA
\textsuperscript{3} The Univ. of Cankiri Karatekin, Faculty of Forestry, Cankiri 18100, Turkey

Correspondence
Yakov Pachepsky, USDA-ARS, Environmental Microbial and Food Safety Lab., Beltsville, MD 20705, USA.
Email: Yakov.Pachepsky@usda.gov

Assigned to Associate Editor Fred Zhang.

Abstract
The selection of the infiltration equation is important in various water research and management fields. Different infiltration equations were found to perform better than others in previous studies. Our objective was to use the Soil Water Infiltration Global (SWIG) database for (a) evaluating and comparing infiltration equations and (b) using the random forest algorithm to investigate how soil properties, land use, and infiltration measurement methods influence the infiltration equation selection. The performance of six equations (Horton, Mezencev, Collis-George, Green–Ampt, Philip, and Swartzendruber) was characterized by the Akaike information criterion obtained after fitting them to 4,830 cumulative infiltration datasets from the SWIG database. Then, the random forest algorithm was applied to predict the best infiltration equation using soil texture class, organic matter content, bulk density, saturated hydraulic conductivity, land use, and infiltration measurement method as inputs. The Horton, Mezencev, and Collis-George models were the best in 36, 24, and 12% of cases, respectively. Swartzendruber, Philip, and Green–Ampt were the best in 11, 10, and 7% cases, respectively. The different error of predicting the best infiltration equation was observed in different parts of input variable space. The infiltration method was by far the most important predictor for a model being the best across the whole database, followed by the organic C content and land use type. Organic C content and land use type were the most important predictors for the tension infiltrometer datasets. The SWIG database presents the opportunity to forecast which infiltration equation will work best in site-specific conditions.

1 | INTRODUCTION

Water infiltration is a fundamental land surface process that controls the hydrological cycle at all scales. The dynamics of infiltration determines water regime and balance at various scales, water availability for runoff and plant consump-

Abbreviations: DT, decision tree; MDA, mean decrease of accuracy; OOB, out-of-bag; RF, random forest; SWIG, Soil Water Infiltration Global.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. Vadose Zone Journal published by Wiley Periodicals LLC on behalf of Soil Science Society of America

https://doi.org/10.1002/vzj2.20125
infiltration rates have to be estimated. Such estimation requires a model of the dependence of infiltration on time, often given in the form of the infiltration equations and dependencies of parameters of this model (these equations) equations on soil and vegetation properties and site-specific conditions water supply rate.

Many equations (not less than 15) were proposed to simulate infiltration (Mishra et al., 2003). In different conditions, different equations performed better. Several studies have compared the accuracy of infiltration equations. Originally, the comparisons were made with data on one or two soils (Davidoff & Selim, 1986; Haverkamp et al., 1988; Lal & Van Doren, 1990; Mbagwu, 1995). Double-ring and single-ring infiltrometers were used in these works. Duan et al. (2011) showed that the Mezencev and Horton infiltration equation performed better than Philip and Kostiakov equations for the data from a double-ring infiltrometer in three soils under lawns. Mishra et al. (2003) analyzed data from 243 experiments in the United States and India and concluded that the semiempirical Singh–Yu general equation performed best compared with other popular infiltration equations. Dastkati et al. (2009) evaluated the Kostiakov, Mezencev, Horton, and Philip infiltration equations with double-ring infiltrometer data from 123 sites in Iran and found that the performance ranking varied across the sites. Shao and Baumgartl (2014) measured infiltration at 28 sites using the rainfall simulator and concluded that the Horton and Holtan equations showed better performance than physically based Green–Ampt and Philip equations. The same infiltration measurement method was used in most of the infiltration equation comparisons. The performance of equations depended on the infiltration measurement method if more than one method was used (Mazloom & Foladmand, 2013).

Pedotransfer functions for infiltration equation coefficients were developed for some regions (e.g., for Green–Ampt equation over the continental United States [Rawls et al., 1983], and for the Philip, Green–Ampt, Kostiakov, and Horton equations over two regions in Brazil [Van de Genachte et al., 1996]). The creation of large international database such as the Soil Water Infiltration Global (SWIG) database (Rahmati, Weihermüller, Vanderborght, et al., 2018; Rahmati, Weihermüller, & Vereecken, 2018) opens the opportunity to create pedotransfer functions for the variety of infiltration equations. To use them in site-specific conditions, one has to decide which infiltration equation should be used in these conditions. Overall, no attempt has been made so far to use site-specific and measurement-specific data to predict which infiltration equation will perform best. Research in this direction becomes feasible with the development of the SWIG database and the availability of powerful algorithms such as random forest (RF) quantifying the complex relationships in databases.

The objective of this work was to use the SWIG database for (a) evaluating and comparing the suitability of infiltration equations and (b) using the RF algorithm to investigate how soil properties, land use, and infiltration measurement methods influence infiltration equation selection.

2 MATERIALS AND METHODS

2.1 The database

Data in SWIG were collected for the 1976–2017 time span. The SWIG database includes 5,023 datasets from 54 countries. It provides cumulative infiltration data, some soil properties (Table 1), land use (Table 2), and infiltration measurement method (Table 3) (Rahmati, Weihermüller, Vanderborght, et al., 2018). Different soil properties have different frequencies of occurrence. Approximately 76% (3,818 out of 5,023) samples have known land use types. Twenty-two categories of land use types in the SWIG database were reclassified into seven categories as presented in Table 2. Agriculture (arable land) is reported as the most frequently reported land use in the SWIG database with 53%, followed by grassland, pasture, garden, forest, urban, and others.

Several infiltration methods were used to measure infiltration rates (Table 3). The disc infiltrometer (disc, mini-disc, micro-disc, hood, and tension infiltrometers) is the most frequently used type of infiltration method in the SWIG database; it was applied to obtain with approximately 51% of datasets. The minidisc infiltrometer is the most frequently reported infiltration measurement method in the SWIG database; it was reported for about 23% of all datasets. The double-ring infiltrometer is the second most frequently applied infiltration measurement method in 16% of datasets. The disc infiltrometer (12% of datasets) and the single-ring infiltrometer (11% of datasets) are ranked as the third and fourth most frequently used infiltration measurement methods, respectively.

The cumulative distribution functions (CDFs) for the length of time of cumulative infiltration from each separated infiltration method are shown in Supplemental Figure S1. Less than
**TABLE 1** Selected statistics of soil properties in Soil Water Infiltration Global (SWIG) datasets used in this work (Rahmati, Weihermüller, Vanderborght, et al., 2018)

| Soil properties          | No. of datasets | Fr | Mean | Min. | Max. | Median | CV % |
|--------------------------|-----------------|----|------|------|------|--------|------|
| Clay, %                  | 3,842           | 76 | 24   | 0    | 80   | 20     | 64   |
| Silt, %                  | 3,842           | 76 | 36   | 0    | 82   | 37     | 52   |
| Sand, %                  | 3,842           | 76 | 41   | 1    | 100  | 38     | 63   |
| Bulk density, g cm⁻³     | 3,295           | 66 | 1.32 | 0.14 | 2.81 | 1.35   | 20   |
| Organic C, %             | 3,102           | 62 | 3    | 0    | 88   | 1      | 200  |

*Fr, frequency.

**TABLE 2** Land use (modified from Rahmati, Weihermüller, Vanderborght, et al., 2018)

| Land use | No. of datasets | Land use | Frequency (no. of datasets) |
|----------|-----------------|----------|------------------------------|
| Agriculture | 2,019           | Forest  | 204                           |
| Grass    | 933             | Urban   | 103                           |
| Pasture  | 233             | Others  | 122                           |
| Garden   | 216             |         |                               |

50 datasets of cumulative infiltration data (microinfiltrometer, hood infiltrometer, linear source method, and point source method) were not included to compute CDF because of the small number of samples. Each of the measurement methods was applied for very different durations, varying mostly from two to three orders of magnitude.

**TABLE 3** Infiltration methods used to measure infiltration (from Rahmati, Weihermüller, Vanderborght, et al., 2018)

| Method                  | No. of datasets |
|-------------------------|-----------------|
| Double ring infiltrometer | 828             |
| Single ring infiltrometer      | 570             |
| Disc infiltrometer           | 607             |
| Minidisc infiltrometer       | 1,140           |
| Microinfiltrometer           | 36              |
| Hood infiltrometer           | 23              |
| Tension infiltrometer        | 752             |
| Guelph permeameter           | 181             |
| Aardvark permeameter         | 50              |
| Rainfall simulator           | 374             |
| Linear source method         | 10              |
| Point source method          | 4               |
| Beerkan(BEST)               | 197             |
| Not reported                | 251             |

**2.2 | Infiltration equations**

Three empirical equations (Horton, Mezencev, and Collis-George) and three theoretical equations (Green–Ampt, Philip, and Swartzendruber) were selected to evaluate and predict their performance in this study. The infiltration equations are shown in Table 4. The software R version 3.53

**TABLE 4** Infiltration equations are included in this study

| Reference        | Infiltration equations |
|------------------|------------------------|
| Horton (1940)    | \( F(t) = f_c t + \frac{f_0}{k} (1 - e^{-kt}) \) |
| Mezencev (1948)  | \( F(t) = k t^a + f_0 \) |
| Collis-George (1977) | \( F(t) = i_0 [\tan(\pi T)]]^{0.5} + K t \) |
| Philip (1957)    | \( F(t) = S t^{0.5} + et \) |
| Green and Ampt (1911) | \( F(t) = K t + \Psi \Delta \theta \ln[1 + \frac{F(t)}{\Psi}] \) |
| Swartzendruber (1987) | \( F(t) = K t + \frac{\Delta \theta}{A} [1 - \exp(-At^{0.5})] \) |

Note. \( F(t) \) = cumulative infiltration at time \( t \) (cm). 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 exponential rate of the approach of the infiltration rate to the final one (h⁻¹). In the Mezencev equation, \( k \) (cm h⁻¹), \( a \) (unitless) and \( f_0 \) (cm h⁻¹) are empirical constants, \( k > 0 \) and \( 0 < a < 1 \). In the Collis-George equation, \( K \) is the infiltration rate at steady state, \( i_0 = (t_c)^{1/2} \), and \( T = t/t_c \); the sorptivity \( S \) is the measurable soil property. In the Philip equation, \( S \) is the sorptivity (cm h⁻¹/²), \( \Psi \) is the soil suction head at the wetting front (cm), and \( \theta \) is the volumetric water content. In the Swartzendruber equation, \( A \) is a parameter depending on soil properties (h⁻⁰.⁵), \( K \) is the saturated hydraulic conductivity (cm h⁻¹), and \( S \) is the soil sorptivity.
FIGURE 1 Decision tree to classify Soil Water Infiltration Global (SWIG) datasets into groups with the same best infiltration equations. Infiltration measurement methods: DI = disk infiltrometer; MI = minidisk infiltrometer; BK = Beerkan; PS = point source; SR = single ring infiltrometer; TI = tension infiltrometer; OC = organic C (%); Collis-George, Green–Ampt, Collis-George, Horton, Mezencev, Philip, and Swartzendruber = infiltration equation names

(R Core Team, 2019) was used to fit equations to datasets using the R `nls_search` routine. There were approximately 200 datasets in which the cumulative infiltration value oscillated with time. Datasets with more than five oscillations were removed. A total of 4,830 cumulative infiltration data at a given time were used for the analysis in this work.

2.3 Application of the RF algorithm

The RF algorithm can model complex relationships between a set of input variables and the output (target) variable. Both input and output variables can be both numerical and categorical. The concept of RF is based on the concept of the decision tree (DT). For a given database, the DT creates the tree-like representation of the relationship between input and output variables. The tree is binary, the data are split into two subgroups, each of those subgroups are split into another two, and so on. For each split, the goal is to create two groups with the most similar target values within the groups and the percentage of all datasets sent to this dataset subgroup. The process ends with having small terminal subgroups. In a DT, each node except terminal nodes represents the test that needs to be performed to send the datasets to one of two branches. A leaf, or terminal node, represents a classification result. The predicted target variable value in these subgroups is found as the average if the target variable is numerical, or by the majority of represented datasets if the target variable is categorical.

An example of the DT with data from the SWIG database is shown in Figure 1. The potential input variables for classifications were the infiltration measurement method, the organic C content, the bulk density, the land use, and the textural class. At the graph’s bottom, there are leaf nodes that present classification results. Each leaf node shows the infiltration equation name, the fractions of datasets classified correctly and misclassified at this node, and the percentage of the total datasets classified to this node. For example, the leftmost node at the bottom represents datasets obtained with disk infiltrometry (DI) and mini-disk infiltrometry (MD). These datasets constitute 13% of the total number of datasets. Fractions of datasets appearing in this group are equal to 0.02 for datasets having the Collis-George equation as the best, 0.00 for the Green–Ampt equation as the best, 0.90 for Horton equation, 0.02 for Mezencev equation, 0.02 for Philip equation, and 0.03 for Swartzendruber equation. This particular subgroup is classified well, and the Horton equation with the fraction of 0.90 dominates. The second from the left leaf node contains a dataset subgroup where the classification results are less certain. The Collis-George equation was voted as the outcome because its fraction of samples was the largest (0.31).
The split conditions are written under classification results at the intermediate nodes. For example, the first split at the top node occurs by the infiltration measurement method. Not all potential classifier variables participate in tree splits—for example, the textural class was not used in this case.

Decision trees work very well with databases that have distinctly different input–output relationships in different database parts. Transparency of the results made DTs a powerful instrument for soil hydraulic database exploration (Lilly et al., 2008; Rawls & Pachepsky, 2002). Decision trees can be used for predictions. In such cases, available data have to be divided into training and testing datasets, and cross-validation has to be performed. Many pairs of training and testing datasets have to be created. Predictions are found by averaging or voting on results from all DTs obtained by training, and the metrics of the prediction accuracy and reliability are found from the testing results. In this way, those metrics can be treated as random values, and this creates an opportunity of using standard statistical procedures for model performance comparison. Pedotransfer relationships have been developed with ensembles of DTs (Gebauer et al., 2020; Jorda et al., 2015).

The RF algorithm is one of several algorithms using DTs for predictions. It builds multiple DTs and merges their results together to get a more accurate and stable prediction. The testing datasets are randomly extracted from the original one, and the training datasets are created by randomly adding data from the original dataset to the positions vacated after the extraction of the testing dataset. This algorithm allows for existence of the same data in training and testing datasets. However, the testing is done only with data in the testing dataset that are not present in the training dataset (i.e., have not been seen during training). Such data are called “out-of-bag,” or OOB, data. The RF algorithm has become popular in pedotransfer work (Araya & Ghezzehei, 2019; Koestel & Jorda, 2014).

The OOB data are used not only for the evaluation of the RF reliability but also for ranking input variables by their importance. A common measure of variable importance is the mean decrease of accuracy, or MDA. After a tree is grown, the OOB samples are passed down the tree, and the prediction accuracy metric is recorded. Then, the values for each input variable are randomly permuted (shuffled) in the OOB samples, and the accuracy metric is again computed. The MDA metric as a result of this permuting is averaged over all trees, normalized by the standard deviation value, and used as a measure of the importance of the permuted variable in the RF (Bhalla, 2019).

The RF algorithm was used to predict the best-performing equation for the dataset’s soil texture class, organic C content, bulk density, land use class, and the infiltration measurement method. The RF built many DTs similar to Figure 1; each tree selects the best model, and the best infiltration equation is the one found in the majority of individual tree outputs (Hastie et al., 2009). The RF algorithm was applied using the randomForest package in R version 3.5.3 (Liaw & Wiener, 2018). Soil properties (soil texture class, organic C contents, and bulk density), land use class, and the infiltration measurement method were used as inputs, and the best infiltration equation name was the output. The “true” output was the name of the equation with the best performance statistic (see below). Before applying the RF classification, datasets were removed if one of the input variables was missing. A total of 1,698 datasets were used for the best equation predicting study. The split of datasets was 75 and 25% for the RF training and testing, respectively. Such splitting was randomly done 500 times to obtain 500 DTs. The number of votes for the correct equation was the accuracy metric.

### 2.4 Performance statistics

The second-order Akaike information criterion, AICc (Burnham & Anderson, 2002), was used to evaluate the performance of the infiltration equations in this study:

\[
AICc = N \ln \left( \frac{\sum_{i=1}^{N} \left[ I_i^{(s)} - I_i^{(m)} \right]^2}{N} \right) + 2(k + 1) + \frac{2k(k + 1)}{N - k - 1}
\]

where \( k \) is the number of estimated parameters in the infiltration equation, \( N \) is the sample size (number of infiltration data points), and \( I_i^{(s)} \) and \( I_i^{(m)} \) are estimated and measured cumulative infiltration at the \( i \)th experimental points, respectively. The smallest value of AICc indicated the best equation.

The accuracy of classification achieved for individual infiltration equations is assessed using the confusion matrix. The element “ij” of this matrix is the number of datasets classified as having the equation \( i \) as the best one, whereas in fact, these datasets have the equation \( j \) as the best one.

The quality of the equation selection achieved with the RF was assessed by the error values that were computed as

\[
\text{Error} = \frac{\text{FP}}{\text{FP} + \text{TP}} \times 100\%
\]

where FP (false positive) is the number of incorrect predictions for an equation to be the best, and TP (true positive) is the number of correct predictions for the equation to be the best.

The overall accuracy was computed from TP, FP, the number of correct predictions for the equation to not be the best (TN, true negative), and the number of incorrect predictions for the equation to not be the best (FN, false negative) (Lu...
et al., 2004):

\[
\text{Overall accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{3}
\]

where TN is correct negative prediction and FN is incorrect negative prediction. The additional statistic to evaluate the classification was the Cohen’s kappa (\(\kappa\)) statistics (Ben-David, 2008):

\[
\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{4}
\]

where \(p_0\) is the observed agreement, and \(p_e\) is the hypothetical probability of change agreement. According to Landis and Koch (1977), less than 0 \(\kappa\) value indicates no agreement, 0–0.20 is slight, 0.21–0.40 is fair, 0.41–0.60 is moderate, 0.61–0.80 is substantial, and 0.81–1 is almost perfect agreement.

3 | RESULTS

3.1 | Evaluating the performance of infiltration equations

All six equations were fitted to the cumulative infiltration time series in each of 4,830 datasets, and AICc values were computed for each fit. The cumulative distribution of the AICc for all models are shown in Supplemental Figure S2. Overall, the distributions were not substantially different. For each dataset, the fitting results of six infiltration equations were ranked using AICc values, and the equation with the smallest AICc was defined as the best (Figure 2). The Horton equation was the best in 1,745 out of 4,830 databases (36%). Mezencev equation (24%) and the Collis-George equation (11.5%) were ranked second and third, respectively. Empirical infiltration equations (Horton, Mezencev, and Collis-George) performed better than theoretical infiltration equations (Philip, Green–Ampt, and Swartzendruber). The Swartzendruber and Philip equations were the most suitable in 535 (11%) and 464 (10%) datasets. The Green–Ampt equation appeared to be the least applicable infiltration equations in this study.

3.2 | Prediction of the best equation and the most influential predictors

The confusion matrix obtained from RF results for the entire dataset is shown in Table 5. The last column provides the error values computed according to Equation 3. The empirical infiltration equations (Horton, Mezencev, and Collis-George) showed prediction accuracy higher than the Philip, Green–Ampt, and Swartzendruber.

The best and the worst performance (35 and 88%) were found for the Horton and Philip infiltration equations, respectively. The difference between training and testing errors was small for the Horton and the Mezencev equations but became more substantial (12 and 19% for the Collis-George and the Green–Ampt equations, respectively). Similar errors in training and testing were found for the Philip and the Swartzendruber equations. The overall accuracy of predictions over all equations was 42%. The \(\kappa\) values for training and testing were within 0.2 and 0.4, indicating a fair agreement.

The input variable ranking by MDA is shown in Figure 3a. The most important predictor was the infiltration measurement method followed by organic C and soil texture class. Land use type and bulk density were not in the top three predictors. Nevertheless, they were relatively close to land use type with MDA in the range from 30 to 40%.

3.3 | Dataset obtained with the same infiltration measurement method

Since the infiltration measurement method appeared to be the leading predictor of the best equation, it was deemed interesting to look at leading predictors for dataset groups that were obtained with the same infiltration measurement method. A large group obtained with the same method was the tension infiltrometer group (476 datasets).

Table 6 shows the confusion matrix obtained from the prediction of the best infiltration model using the tension infiltrometer dataset. The Collis-George infiltration equation had the best error with training data, and its errors with the testing data were much better than errors of other models. The overall accuracy across all equations was 27%, which was much lower than for the entire dataset in Table 5. The \(\kappa\) values for training and testing were 0.086 and 0.066, which were also lower than for the entire dataset.
TABLE 5  Confusion matrix of the predictions obtained with all data for the Horton (H), Mezencev (M), Collis-George (C), Philip (P), Swartzendruber (S), and Green–Ampt (G) equations

| Predicted best equation | Actual best equation | Error |
|-------------------------|----------------------|-------|
|                         | H                   |       |
|                         | M                   |       |
|                         | C                   |       |
|                         | P                   |       |
|                         | S                   |       |
|                         | G                   |       |
| H                       | 314 (108)           | 34.9 (32.9) |
| M                       | 81 (21)             | 53.0 (50.6) |
| C                       | 49 (21)             | 54.2 (66.6) |
| P                       | 32 (17)             | 87.8 (90.6) |
| S                       | 47 (15)             | 86.0 (85.1) |
| G                       | 31 (13)             | 78.6 (97.5) |

*Values without parentheses are for training datasets, values in parenthesis are for testing datasets. \( \kappa = 0.265 (0.223) \).

FIGURE 3  Importance of best model predictors assessed by the mean decrease of accuracy: (a) the entire dataset, (b) dataset obtained with the tension infiltrometry

The rankings of the best equation predictors were different for the entire dataset and tension infiltrometer datasets. The organic C content was the most important predictor in the tension infiltrometer dataset. Land use and bulk density were equally important (Figure 3b).

4  | DISCUSSION

The comparison of six infiltration equations by the AICc criterion showed that empirical infiltration equations provided a better fit compared with theoretical infiltration equations.
Past research provided both similar and contradicting to our conclusions. The Horton equation was found to be the most suitable infiltration equation when compared with the Kostiakov, the modified-Kostiakov, and the Philip equations (Ogbe et al., 2011). Previous research showed that Horton and Mezencev infiltration equations performed better than Philip and Kostiakov (Duan et al., 2011). Shao and Baumgartl (2014) demonstrated that empirical infiltration equations were more accurate than theoretical infiltration equations when Philip, Green–Ampt, Horton, and Holtan equations were compared. Mishra et al. (2003) found that theoretical infiltration equations performed better in a study for evaluation of 14 popular infiltration equations based on the Nash–Sutcliffe efficiency criterion using 243 sets of infiltration test data of various soil types from India and the United States. Jha et al. (2019) concluded that the theoretical infiltration equation (Swartzendruber) provided the best reliable infiltration equation by comparing five infiltration equations using 72 infiltration data. The above studies were conducted with relatively small databases with soil samples from specific and limited locations compared with the SWIG database. The evaluation of suitability for infiltration equations using the SWIG database might result in more broadly applicable conclusions. In general, there is no reason for theoretical equations to be more accurate than empirical equations. The pore space of field soils is more complex than the pore space assumed in theoretical equation development. This study was limited to six infiltration equations, further research is possible to select the most suitable infiltration equations by including other infiltration equations from the literature.

It is also worth mentioning that the Horton and Mezencev equations were not only the most accurate but also were the most accurately predicted as the best equations (Table 5). Besides, the prediction error was very close in training and testing datasets. That may mean that the predictor set includes the major determinants for the selection of the best equation.

The infiltration method was the most influential predictor for the selection of the most suitable infiltration equation. There can be several reasons for that, including the scale effect caused by differences in the contact area between infiltrometer and soil. The contact area for the minidisc infiltrometer and double-ring infiltrometer is approximately 16 and 700 cm\(^2\), respectively. These differences can make measurements sensitive to different levels of soil heterogeneity (Pachepsky et al., 2014). Rahmati, Weihermüller, Vanderborght, et al. (2018) noted that different infiltration equations should be used for measurements made with different methods. Prediction of the infiltration equations can have a different efficiency for different infiltration measurement methods.

When percentages of sand, silt, and clay were used instead of soil texture class (the results are shown as supporting information in Supplemental Tables S1 and S2), the overall accuracy was 44.8 and 34.1%, respectively. The overall accuracy was not much different from overall accuracy when computing RF input soil texture class (suggested by the anonymous reviewer). Given the importance of infiltration equations in land surface modeling (Vereecken et al., 2019), the efficiency of the soil texture class as a predictor of infiltration equation suitability makes possible large-scale maps to inform land surface modeling on selection of the infiltration equation.

The organic C content was found to be one of the important predictors (Figure 3). This happens because soil organic C contents highly are related to soil structure (Haynes & Beare, 1996; Huntington, 2007). The soil bulk density was not found as one of the most important predictors in the results from the entire dataset and from the tension infiltrometer data. The bulk density can reflect different aspects of soil structure caused by compaction (Hakansson & Lipiec, 2000) or aggregation (Aksakal et al., 2019) in the soil. It is also possible that the land use provides enough information for predictions of the best equation and makes the bulk density less critical predictor. Several previous studies stressed out the land use type impact on hydraulic conductivity (Kidwell et al., 1997; Shao & Baumgartl, 2014).

The RF input values that were conducive for the Horton equation to be the best provided a moderately accurate prediction across the whole database. Other input values pointed to other infiltration equations as the best ones, but the error of these predictions was high. Similarly, in measurements with the tension infiltrometer, the RF input values indicated that the Collis-George equation was the best led to the moderate accuracy in predictions. In contrast, low accuracy was found with the input sets pointing to other equations as the best ones. These results of predicting the best model appear to demonstrate a common, but not always analyzed, feature in pedotransfer work—different accuracy of pedotransfer in different parts of the input domain. For example, in studies of scaled pedotransfer functions, Amanabadi et al. (2019) concluded that predictions of water retention curves of fine-textured soil significantly better than that of the coarse-textured soils. Grote and Leverett (2019) compared pedotransfer functions for high-resolution mapping of hydraulic conductivity using GPR. They observed that the GPR-based estimates were more accurate in the sandy loam than in the silty clay loam. Why the ability of some equations to be the best can be predicted better than for other equations presents an interesting research question.

Additional predictors of infiltration equation suitability may be found and used. The initial water content in the soil should be examined as a predictor because it can influence the infiltration rate which was considered for predicting hydraulic conductivity (Araya & Ghezzehei, 2019). The initial water content in the soil was not used in this study, since it was only available for 31% of infiltration data in the SWIG database. Furman et al. (2006) showed that initial and boundary conditions may be influential factors in infiltration equation.
performance. One may argue that some infiltration equations may be better suited for one-dimensional flow than for three-dimensional and vice versa, and that might be incorporated in the analysis. There are two aspects to that. First, some equations are known to work well for both geometries. In particular, the Philip equation (Table 4) was derived from the one-dimensional water flow equation (Philip, 1957). However, Havercamp et al. (1994) showed that the same equation applies to the three-dimensional infiltration in disc infiltrometry for short and medium times. This was confirmed in recent experiments of Kargas et al. (2018), where the linear dependence on time was demonstrated for the difference between three- and one-dimensional cumulative infiltration. Similarly, the Mezencev equation has performed well with data from both one- and three-dimensional flow (Kumar & Sihag, 2019). Second, the performance of an infiltration equation also depends on the stage of the infiltration that has been reached in measurements. In measurements that did not reach the stationary stage, the presumably one-dimensional equations such as Horton’s or Swartzendruber’s may perform better than the presumably three-dimensional equations for some pore space geometries. The SWIG database includes both relatively short and relatively long measurements for all measurement methods (Supplemental Figure S1). Including the infiltration stage in the analysis might help create a better prediction of the most suitable equation. This can probably be done for small subsets of the SWIG database and will require an algorithm to decide at which stage the infiltration measurements ended. This needs in-depth research, given the observation of Warrick (1992) who researched the equations for disk infiltrometers and pointed out that the approach to the steady-state solution may take considerably longer than what is commonly reported in the literature for field applications. Nevertheless, predicting which equation will be suitable at a particular infiltration stage presents an interesting research avenue.

The analysis is done in this work for tension infiltrometer data infiltration measurement methods presents an example. It shows that, depending on the measurement method more appropriate for the project at hand, the SWIG database allows one to select a subset of data with the soil and landscape properties that are available for this project, and evaluate the precision of the selection of the infiltration equation.

5 | CONCLUSIONS

We evaluated which infiltration equation is more suitable using RF classification and investigate how soil properties, land use, and infiltration methods control infiltration equations. We used 4,830 cumulative infiltration data from the SWIG database. The Horton equation was found as the most suitable equation and Mezencev and Collis-George equations were ranked second and third, respectively. The Green–Ampt equation was found as the least suitable infiltration equation. For the entire dataset, the most important predictor was the infiltration measurement method followed by organic C and land use type. The ranking of predictors for selection the most suitable infiltration equation was different for different infiltration measurement methods. The accuracy of predicting the best model appeared to be different for different parts of the input domain. The SWIG database presents an important resource for selecting the appropriate infiltration equation for project-specific infiltration methods and available predictors.

ACKNOWLEDGMENTS

Dr. Gülay Karahan thanks the Scientific and Technological Council of Turkey (TUBITAK) for the financial support given under the grant of BIDEP (2219).

AUTHOR CONTRIBUTIONS

Seongyun Kim: Investigation; Writing-original draft. Manan Sharma: Funding acquisition; Resources; Supervision; Writing-review & editing. Gülay Karahan: Data curation; Validation; Writing-review & editing. Yakov Pachepsky: Conceptualization; Methodology; Validation; Writing-review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Seongyun Kim https://orcid.org/0000-0002-0196-5037
Gülay Karahan https://orcid.org/0000-0003-1285-6546
Manan Sharma https://orcid.org/0000-0002-8585-0308
Yakov Pachepsky https://orcid.org/0000-0003-0232-6090

REFERENCES

Aksakal, E. L., Barik, K., Angin, I., Sari, S., & Islam, K. R. (2019). Spatio-temporal variability in physical properties of different textured soils under similar management and semi-arid climatic conditions. Catena, 172, 528–546. https://doi.org/10.1016/j.catena.2018.09.017
Amanabadi, S., Vazirinia, M., Vereecken, H., Vakilian, K. A., & Mohammadi, M. H. (2019). Comparative study of statistical, numerical and machine learning-based pedotransfer functions of water retention curve with particle size distribution data. Eurasian Soil Science, 52, 1555–1571. https://doi.org/10.1134/S106422931930001X
Araya, S. N., & Ghezzehei, T. A. (2019). Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. Water Resources Research, 55, 5715–5737. https://doi.org/10.1029/2018WR024357
Ben-David, A. (2008). Comparison of classification accuracy using Cohen’s weighted kappa. Expert Systems with Applications, 34, 825–832. https://doi.org/10.1016/j.eswa.2006.10.022
Bhalla, D. (2019). A complete guide to random forest in R. Listendata. https://www.listendata.com/2014/11/random-forest-with-r.html#Important-Features-Variable-Importance
Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.), Springer.

Collis-George, N. (1977). Infiltration equations for simple soil systems. *Water Resources Research, 13*, 395–403. https://doi.org/10.1029/WR013i002p00395

Dastkhi, S. G., Homaei, M., Mahdian, M. H., & Kouchakzadeh, M. (2009). Site-dependence performance of infiltration models. *Agricultural Water Management, 23*, 2777–2790.

Davidoff, B., & Selim, H. M. (1986). Goodness of fit for eight water infiltration models. *Soil Science Society of America Journal, 50*, 759–764. https://doi.org/10.2136/sssaj1986.03615995005000030039x

Duan, R., Fedler, C. B., & Borrelli, J. (2011). Field evaluation of infiltration models in lawn soils. *Irrigation Science, 29*, 379–389. https://doi.org/10.1007/s00271-010-0248-y

Furman, A., Warrick, A. W., Zerihun, D., & Sanchez, C. A. (2006). Modified Kostiakov infiltration function: Accounting for initial and boundary conditions. *Journal of Irrigation, 132*, 587–596. https://doi.org/10.1061/(ASCE)0733-9437(2006)132:6(587)

Gebauer, A., Ellinger, M., Brito, M., Gomez, V., & Ließ, M. (2020). Development of pedotransfer functions for water retention in tropical mountain soil landscapes: Spotlight on parameter tuning in machine learning. *Soil, 6*, 215–229. https://doi.org/10.1914/sol-6-215-2020

Green, W. H., & Ampt, C. A. (1911). Studies on soil physics, I. Flow of air and water through soils. *Journal of Agricultural Science, 4*, 1–24.

Grote, K., & Leverett, K. (2019). Comparison of pedotransfer functions for high-resolution mapping of hydraulic conductivity in agricultural soils using GPR. In *AGU Fall Meeting Abstracts*, (Vol. 2019, NS31A-0762). American Geophysical Union.

Hakansson, I., & Lipiec, J. (2000). A review of the usefulness of relative bulk density values in studies of soil structure and compaction. *Soil and Tillage Research, 53*, 71–85. https://doi.org/10.1016/S0167-1987(99)00095-1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). Springer. http://www.springerlink.com/index/D7x7KX6772HQ2135.pdf

Haverkamp, R., Kutilek, M., Parlane, Y., Rendon, L., & Krejca, M. (1988). Infiltration under ponded conditions: 2. Infiltration equations tested for parameter time-dependence and predictive use. *Soil Science, 145*, 317–329. https://doi.org/10.1007/90010694-198805000-00001

Haverkamp, R., Ross, P. J., Smettem, K. R. J., & Parlange, J. Y. (1994). Development of infiltration models for water retention in tropical mountain soil landscapes: Spotlight on parameter tuning in machine learning. *Soil, 6*, 215–229. https://doi.org/10.1914/sol-6-215-2020

Haynes, R. J., & Beare, M. H. (1996). Aggregation and organic matter storage in mesothermal, humid soils. In M. R. Carter & B. A. Stewart (Eds.), *Structure and soil organic matter storage in agricultural soils* (pp. 213–262). CRC Press.

Horton, R. E. (1940). An approach towards a physical interpretation of infiltration capacity. *Soil Science Society of America Journal, 5*, 399–417. https://doi.org/10.2136/sssaj1941.03615995005000000075x

Huntington, T. G. (2007). Available water capacity and soil organic matter. In *Encyclopedia of soil science* (2nd ed., pp. 139–143). Taylor and Francis. https://doi.org/10.1081/E-ESS-120018496

Jha, M. K., Mahapatra, S., Mohan, C., & Pohshna, C. (2019). Infiltration characteristics of lateritic vadose zones: Field experiments and modeling. *Soil and Tillage Research, 187*, 219–234. https://doi.org/10.1016/j.still.2018.12.007

Jorda, H., Bechtold, M., Jarvis, N., & Koestel, J. (2015). Using boosted regression trees to explore key factors controlling saturated and near-saturated hydraulic conductivity. *European Journal of Soil Science, 66*, 744–756. https://doi.org/10.1111/ejss.12249

Kargas, G., Londra, P., Anastasiou, K., & Kerkides, P. (2018). A note on one-and three-dimensional infiltration analysis from a mini disc infiltrometer. *Water, 10*, 1783. https://doi.org/10.3390/w10121783

Kidwell, M. R., Weltz, M. A., & Guertin, D. P. (1997). Estimation of Green-Ampt effective hydraulic conductivity for rangelands. *Rangeland Ecology & Management/ Journal of Range Management, 50*, 290–299.

Koestel, J., & Jorda, H. (2014). What determines the strength of preferential transport in undisturbed soil under steady-state flow? *Geoderma, 217–218*, 144–160.

Kumar, M., & Sihag, P. (2019). Assessment of infiltration rate of soil using empirical and machine learning-based models. *Irrigation and Drainage, 68*, 588–601. https://doi.org/10.1002/ird.2332

Lal, R., & Van Doren, J., D. M. (1990). Influence of 25 years of continuous corn production by three tillage methods on water infiltration for two soils in Ohio. *Soil and Tillage Research, 16*, 71–84. https://doi.org/10.1016/10167-1987(90)00226-6

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

Liav, A., & Wiener, M. (2018). *Breiman and Cutler’s random forests for classification and regression*. R Package 'randomForest. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/randomForest/randomForest/randomForest.pdf

Lilly, A., Nemes, A., Rawls, W. J., & Pachepsky, Y. A. (2008). Probabilistic approach to the identification of input variables to estimate hydraulic conductivity. *Soil Science Society of America Journal, 72*, 16–24. https://doi.org/10.2136/sssaj2006.0091

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D. S., Poulin, B., Anvik, J., Macdonell, C., & Eisen, R. (2004). Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics, 20*, 547–556. https://doi.org/10.1093/bioinformatics/btg447

Mazloom, H., & Foldmand, H. (2013). Evaluation and determination of the coefficients of infiltration models in Marvdasht region, Fars province. *International Journal of Advanced Biological and Biomedical Research, 1*, 822–829.

Mbagwu, J. S. C. (1995). Testing the goodness of fit of infiltration models on a sandy soil in Lafia, Southern Guinea Savanna Zone of Nigeria. *Production Agriculture and Technology Journal, 7*, 116–126.

Pachepsky, Y. A., Guber, A. K., Yakirevich, A. M., McKee, L., Cady, R. E., & Nicholson, T. J. (2014). Scaling and pedotransfer in numerical simulations of flow and transport in soils. *Vadose Zone Journal, 13*(12). https://doi.org/10.2136/vzj2014.02.0020
Philip, J. R. (1957). The theory of infiltration: 4. Sorptivity and algebraic infiltration equations. Soil Science, 84, 257–264. https://doi.org/10.1097/00010694-195709000-00010

R Core Team. (2019). The R project for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/

Rahmati, M., Weihermüller, L., Vanderborght, J., Pachepsky, Y. A., & Vereecken, H. (2018). Development and analysis of the Soil Water Infiltration Global database. Earth System Science Data, 10, 1237–1263. https://doi.org/10.5194/essd-10-1237-2018

Rahmati, M., Weihermüller, L., & Vereecken, H. (2018). Soil Water Infiltration Global (SWIG) database. Supplement to Rahmati, M et al. (2018): Development and Analysis of Soil Water Infiltration Global Database. PANGAEA. https://doi.org/10.1594/PANGAEA

Rawls, W. J., Brakensiek, D. L., & Miller, N. (1983). Green-Ampt infiltration parameters from soils data. Journal of Hydraulic Engineering, 109, 62–70. https://doi.org/10.1061/(ASCE)0733-9429(1983)109:1(62)

Rawls, W. J., & Pachepsky, Y. A. (2002). Soil consistence and structure as predictors of water retention. Soil Science Society of America Journal, 66, 1115–1126. https://doi.org/10.2136/sssaj2002.1115

Shao, Q., & Baumgartl, T. (2014). Estimating input parameters for four infiltration models from basic soil, vegetation, and rainfall properties. Soil Science Society of America Journal, 78, 1507–1521. https://doi.org/10.2136/sssaj2014.04.0122

Smith, R., Smettem, K., Broadbridge, P., & Woolhiser, D. A. (2002). Infiltration theory for hydrologic applications (Water Resources Monograph, Vol. 15). American Geophysical Union.

Swartzendruber, D. (1987). A quasi-solution of Richards equation for the downward infiltration of water into soil. Water Resources Research, 23, 809–817. https://doi.org/10.1029/WR023i005p00809

Thompson, S. A. (2017). Hydrology for water management. CRC Press.

Van de Genachte, G., Mallants, D., Ramos, J., Deckers, J. A., & Feyen, J. (1996). Estimating infiltration parameters from basic soil properties. Hydrological Processes, 10, 687–701. https://doi.org/10.1002/(SICI)1099-1085(199605)10:5<687::AID-HYP311>3.0.CO;2-P

Vereecken, H., Weihermüller, L., Assouline, S., Šimůnek, J., Verhoef, A., Herbst, M., Archer, N., Mohanty, B., Montzka, C., Vanderborght, J., Balsamo, G., Bechtold, M., Boone, A., Chadburn, S., Cuntz, M., Decharme, B., Ducharne, A., Ek, M., Garrigues, S., ..., & Xue, Y. (2019). Infiltration from the pedon to global grid scales: An overview and outlook for land surface modeling. Vadose Zone Journal, 18(1). https://doi.org/10.2136/vzj2018.10.0191

Warrick, A. W. (1992). Models for disc infiltrometers. Water Resources Research, 28, 1319–1327. https://doi.org/10.1029/92WR00149

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kim S, Karahan G, Sharma M, & Pachepsky Y. The site-specific selection of the infiltration model based on the global dataset and random forest algorithm. Vadose Zone J. (2021);20:e20125. https://doi.org/10.1002/vzj2.20125