Published November 7, 2019

Original Research

Core Ideas

- PTFs for water contents at specific pressure heads were developed.
- Covariate shift increased uncertainty in PTF predictions.
- Relative importance of predictors in machine learning PTFs was determined.

Development and Uncertainty Assessment of Pedotransfer Functions for Predicting Water Contents at Specific Pressure Heads

Ali Mehmandoost Kotlar,* Quirijn de Jong van Lier, Alexandre Hugo C. Barros, Bo V. Iversen, and Harry Vereecken

There has been much effort to improve the performance of pedotransfer functions (PTFs) using intelligent algorithms, but the issue of covariate shift, i.e., different probability distributions in training and testing datasets, and its impact on prediction uncertainty of PTFs has been rarely addressed. The common practice in PTF generation is to randomly separate the dataset into training and testing subsets, and the outcomes of this random selection may be different if the process is subject to covariate shift. We evaluated the impact of covariate shift generated by data shuffling and detected by Kolmogorov–Smirnov test for the prediction of water contents using soil databases from Denmark and Brazil. The soil water contents at different pressure heads were predicted by developing linear and stepwise regression besides machine learning based PTFs including Gaussian process regression and ensemble method. Regression based PTFs for the Brazilian dataset resulted in better predictions compared with machine learning methods, which in their turn estimated high water contents in Danish soils more accurately. One hundred PTFs were developed for water content at specific pressure heads by data shuffling. From these, 100 sets of fitted van Genuchten parameters were obtained representing the generated uncertainty. Data shuffling led to covariate shift, resulting in uncertainty in water content prediction by the PTFs. Inherent variability of data may lead to increased prediction uncertainty. For correlated data, simple regression models performed as good as sophisticated machine learning methods. Using PTF-predicted water contents for van Genuchten retention parameter fitting may lead to a high uncertainty.

Abbreviations: BD, bulk density; ENS, ensemble regression with bagging aggregation; GP, Gaussian process; LM, linear model; OM, organic matter; PTF, pedotransfer function; SLM, stepwise linear model.

A.M. Kotlar and Q. de Jong van Lier, Centre for Nuclear Energy in Agriculture (CENA/USP), Univ. of São Paulo, Caixa Postal 96, 13416-903 Piracicaba (SP), Brazil; A.H.C. Barros, Brazilian Agricultural Research Corporation (EMBRAPA), 51020-240 Recife (PE), Brazil; B.V. Iversen, Dep. of Agroecology, Aarhus Univ., Blichers Allé 20, 8830 Tjele, Denmark; H. Vereecken, Institute of Bio- and Geosciences (IBG-3), Agrosphere, Forschungszentrum Jülich, 52425, Jülich, Germany. *Corresponding author (aliko@usp.br).

Received 24 June 2019.
Accepted 11 Oct. 2019.

Citation: Kotlar, A.M., Q. de Jong van Lier, A.H.C. Barros, B.V. Iversen, and H. Vereecken. 2019. Development and uncertainty assessment of pedotransfer functions for predicting water contents at specific pressure heads. Vadose Zone J. 18:190063. doi:10.2136/vzj2019.06.0063

Pedotransfer functions (PTFs) correlate more readily available soil characteristics such as texture, particle size fractions, organic matter (OM), and bulk density (BD) to properties that are more difficult to measure (Bouma, 1989). Pedotransfer functions to predict soil hydraulic properties, especially soil water retention, are among the most frequently used. These functions are used in the simulation of soil processes across scales and in land surface and Earth system models and may be an interesting alternative to direct measurements (McBratney et al., 2002; Van Looy et al., 2017). Although these PTF-based indirect estimations significantly reduce experimental cost and time, they introduce uncertainty in simulations of soil processes.

Pedotransfer functions are developed using some kind of statistical fitting procedure. Regression-based PTFs for the prediction of water content θ as a function of pressure head h using particle size fractions, OM content, and BD were developed initially by Gupta and Larson (1979) and Rawls et al. (1982) using data from the United States. Similar functions were developed and tested later by Minasny et al. (1999), Tomasella et al. (2000), and Børgesen et al. (2008) for soils from New Zealand, Brazil, and Denmark, respectively. Further developments beyond the classical regression analysis include techniques like artificial neural networks (Schaap and Leij, 1998a, 1998b; Minasny and McBratney,

2002; Merdun et al., 2006; Baker and Ellison, 2008; Campos de Oliveira et al., 2017; D'Emilio et al., 2018), pattern recognition based methods including support vector machine (Nemes et al., 2006; Lamorski et al., 2008; Khlosi et al., 2016), Gaussian process regression (Kotlar et al., 2019b), and the ensemble approach (Baker and Ellison, 2008; Cichota et al., 2013; Liao et al., 2015), all of which contributed to the improvement of PTF performance in predicting soil hydraulic properties.

Based on a measured dataset and statistically derived, the uncertainty of PTFs when extrapolated to other regions, climates, or soil types is hard to assess. Several studies (Schaap and Leij, 1998a; Guber et al., 2006; Pachepsky and Rawls, 2004; Patil and Singh, 2016) showed that PTFs are strongly dependent on location, and it has therefore been recommended to apply PTFs calibrated on small local datasets rather than using PTFs that were developed on datasets from other regions. There is still need for concern regarding the inability of PTFs to be extrapolated (Tranter et al., 2009). Nevertheless, this kind of extrapolation is common practice in regions or countries with a lack of measured soil hydraulic data (Nemes et al., 2009), using widely accepted PTFs such as Rosetta (Schaap et al., 2001).

The quantification of uncertainty in PTFs is important, especially when the output is used in water balance simulations (McBratney et al., 2002; Deng et al., 2009). According to the principles given by McBratney et al. (2002), PTF uncertainty should be quantified and the PTF with minimum variance should be used among the available sets of PTFs for a specific target. Previous studies have addressed the issue of uncertainty in PTFs such as using bootstrapping (Efron and Tibshirani, 1994) to analyze modeling uncertainty when developing PTFs (Schaap and Leij, 1998b; Ye et al., 2007) and evaluating the uncertainty due to measurement errors in input variables (Minasny et al., 1999; Chirico et al., 2010). The issue of uncertainty in PTFs becomes more important for those regions where data for PTF development are rare and globally accepted PTFs are applied to obtain hydraulic parameters when the probable outcome uncertainty would be propagated into hydrological models. Given protocols to determine the similarity between calibrated data and the target of interest with PTF developments by Tranter et al. (2009, 2010) for cation exchange capacity and the wilting point were based on metric distance and fuzzy *k*-means to remove such barriers.

A common assumption in the development of PTFs using supervised learning algorithms is that data for training and testing present the same statistical properties and frequency distribution. In practice, however, this may not be the case and a single probability distribution function is not capable of describing the collected data. The difference between the frequency distribution of the training and testing datasets, called *covariate shift* (Sugiyama, 2012), may lead to biased models with a low generalizability of the results (Bishop, 2006; Chung et al., 2018).

To assess the uncertainty of PTFs under covariate shift, the whole dataset can be rearranged in advance using Monte Carlo shuffling to select training and testing subsets (Schaap and Leij, 1998b). Pedotransfer functions can then be developed on each shuffled dataset and analyzed in terms of the uncertainty of their prediction capability. This random sampling of training and testing datasets has been addressed in a few studies such as Zhao et al. (2016, 2017), who demonstrated a higher uncertainty in artificial neural network PTFs for the prediction of saturated hydraulic conductivity as shown by highly scattered RMSE values obtained from 400 random samplings. Similarly, Jarvis et al. (2013) found a significant range of the bootstrapped normalized model coefficients for the prediction of saturated hydraulic conductivity.

The general objective of this study was to contribute to the evaluation of PTF development techniques and their performance. To do so, our specific objectives were (i) to develop and compare simple regression and complex machine learning based techniques to develop a PTF-$\theta(h)$ for the prediction of water content at specific pressure heads for soils from Denmark and Brazil, (ii) assess the relative importance of predictors in machine learning based PTFs if machine learning methods perform better than simple methods, (iii) investigate the uncertainty of PTFs under the effect of differently distributed training and testing datasets, and (iv) assess the effect of uncertainty in estimated water contents on fitted van Genuchten (1980) parameters.

Material and Methods

Soil Datasets

Data from samples collected in the northeastern region of Brazil (Brazil-NE, Fig. 1a) were retrieved from Barros et al. (2013) and a dataset of Embrapa, the Brazilian Agricultural Research Corporation. The database comprised 838 samples with water

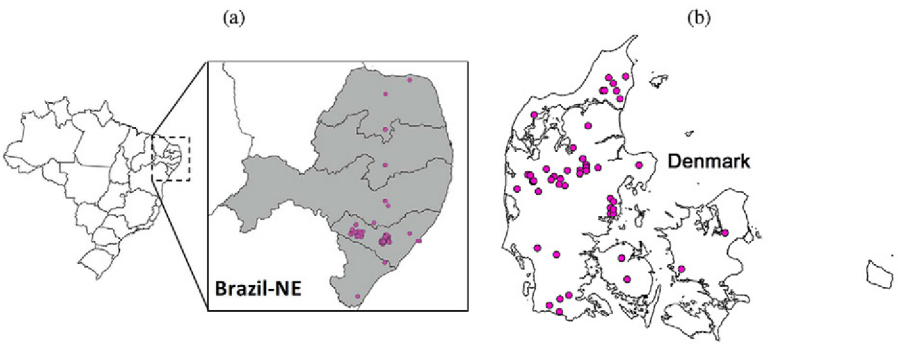

Fig. 1. Location of sample sites for water retention (Brazil-NE and Denmark datasets).

contents determined at −0.6, −1, −3, −5, −10, −20, and −150 m pressure head ($\theta_{0.6}$, θ_1, θ_3, θ_5, θ_{10}, θ_{20}, θ_{150}, respectively), particle size fractions (sand, silt, and clay contents), OM, and BD.

Soil data from a temperate region were obtained from a Danish database containing 186 samples (Fig. 1b), mostly collected from the Jutland peninsula, western Denmark (Iversen et al., 2011; Børgesen and Schaap, 2005), containing similar information but with water contents at only four pressure heads: −0.1, −1, −10, and −150 m ($\theta_{0.1}$, θ_1, θ_{10}, and θ_{150}, respectively). In both datasets, particle size was classified according to the USDA Soil Taxonomy as clay (<0.002 mm), silt (0.002–0.05 mm), and sand (0.05–2 mm). The saturated water content (θ_s) was considered equal to the total porosity and was calculated from BD and particle density (assumed equal to 2.65 g cm^{-3}).

Model Description

Covariate Shift and Development of Pedotransfer Functions

For each dataset (Denmark and Brazil-NE), regression and machine learning based PTFs were developed for estimating water contents at the available pressure heads. The training dataset comprised 70% of data randomly allocated; the remaining 30% were the testing dataset. The presence of covariate shift, when the training and testing dataset have different statistical distributions, was tested by randomly shuffling the data and reallocating them 100 times among the training and testing sets. For each random shuffle, a PTF for each pressure head was developed based on the resulting training set. The statistical properties of the training and testing stages were compared for each shuffle by graphically representing the correlation between mean and variance values of the training and testing data and by performing a Kolmogorov–Smirnov test to explore whether training and testing datasets presented the same frequency distributions.

Regression methods used to develop PTFs included two more common simple methods (the linear model [LM] and the stepwise linear model [SLM]), and two more complex machine learning methods: Gaussian process (GP) regression and ensemble regression with bagging aggregation (ENS).

In the LM, all predictor variables are fitted at once. The SLM adds predictors one by one and computes the p value of an F statistic to add or remove potential variables; the final model is obtained when no single step improves the model.

In Gaussian process (GP) regression, nearest neighbors are used by considering the distance between them based on a covariance (or kernel) function. The closeness or similarity between two points (distance) is given by kernel functions (Rasmussen and Williams, 2006). Kernel similarities between a test point and each point of the training data are found to predict the target of the test point, thus kernel values of far-away points tend to zero (Kotlar et al., 2019b). In mathematical form, GP regression can be represented as

$$\begin{bmatrix} Y_{tr} \\ Y_{ts} \end{bmatrix} = \mathrm{GP}\left(0, \begin{bmatrix} K_{tr} & K_{trs} \\ K_{trs}^{T} & K_{ts} \end{bmatrix}\right) \quad [1]$$

where Y_{tr} and Y_{ts} are training and testing targets (e.g., water contents) and K_{tr}, K_{ts}, and K_{trs} are the covariance of the training data, the testing data, and the covariance between the training and testing data, respectively. Considering a Gaussian likelihood function, the predictive mean y_{ts} for a given test point (x_{ts}) is

$$y_{ts} = \mathbf{K}_{x_{ts}}^{T} K_{tr}^{-1} Y_{tr} \quad [2]$$

where $\mathbf{K}_{x_{ts}}^{T}$ is the vector with the distances from x_{ts} to each training point. The optimization of kernel parameters and other details are given in Kotlar et al. (2019b, 2019c), who successfully applied Gaussian regression. The length scale of each predictor extracted from its squared exponential kernel function shows the weight or importance of the respective predictor in the prediction by a GP PTF. The relative importance of each predictor is computed by dividing each predictor importance by the sum of the importance of all the predictors.

The ENS is based on the aggregation of results from multiple learning algorithms (decision tree or weak learners) into a robust ensemble predictor (Zhang and Ma, 2012). The bootstrap aggregation (bagging) algorithm generally forms deep trees with less concern about overfitting (Møller et al., 2018). The relative importance of each predictor in random forest is obtained by summing the changes in the errors due to each split and dividing the sum by the number of branch nodes.

Water Retention Fitting

After obtaining 100 PTFs for each development technique as described above, water contents $\theta_{0.6}$, θ_1, θ_3, θ_5, θ_{10}, θ_{20} and θ_{150} for Brazil-NE and $\theta_{0.1}$, θ_1, θ_{10} and θ_{150} for Denmark, corresponding to the respective pressure heads, were obtained using the best of the developed PTFs among the LM, SLM, GP, and ENS methods. Water content predictions were fitted to the van Genuchten (1980) equation, using the RETC software (van Genuchten et al., 1991):

$$S_e(h) = \frac{\theta(h) - \theta_r}{\theta_s - \theta_r} = \left(1 + |\alpha h|^n\right)^{-m} \quad [3]$$

where S_e is the effective saturation, θ_s and θ_r are saturated and residual volumetric water contents, respectively, α (m^{-1}) is a scale parameter, and m and n are curve shape parameters, with $m = 1 - 1/n$.

Model Evaluation

The performance of the developed PTFs for predicting the target (water content) was evaluated by the root mean square error (RMSE), the coefficient of determination R^2, representing the proportion of the variance in the measured data, and finally the Nash–Sutcliffe efficiency (NSE), showing the match between observed and predicted values:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{p=1}^{n}\left(Y_{est,p} - Y_{obs,p}\right)^2}{n}} \quad [4]$$

$$R^2 = \left[\frac{\sum_{p=1}^{n} \left(Y_{\text{obs},p} - \overline{Y}_{\text{obs},p}\right)\left(Y_{\text{est},p} - \overline{Y}_{\text{est},p}\right)}{\sqrt{\sum_{p=1}^{n} \left(Y_{obs,p} - \overline{Y}_{\text{obs},p}\right)^2 \sum_{p=1}^{n} \left(Y_{est,p} - \overline{Y}_{\text{est},p}\right)^2}} \right]^2 \quad [5]$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n} \left(Y_{\text{est},p} - Y_{\text{obs},p}\right)^2}{\sum_{i=1}^{n} \left(Y_{\text{obs},p} - \overline{Y}_{\text{obs},p}\right)^2} \quad [6]$$

where Y_{obs} and Y_{est} are the measured and PTF-predicted target variables (water contents) and $\overline{Y}_{\text{obs}}$ and $\overline{Y}_{\text{est}}$ are the average values of the corresponding variables (Krause et al., 2005). Uncertainty in the PTF performance was assessed by considering the distribution of the RMSE obtained from the PTFs developed for each of the 100 data shuffles.

Results and Discussion

Data Analysis

The soils in the Brazilian and Danish datasets are shown in Fig. 2 on a USDA textural triangle. The Danish soils are mainly sandy and sandy loam (65% of the samples), and the majority of the remaining samples are loamy sands (Fig. 2). The Brazil-NE dataset is about 80% sandy clay loam and sandy loam soil samples. The remaining 20% mainly consists of loamy sand and clay loam samples (Fig. 2).

The descriptive statistics of all soil properties used in PTF development are shown in Table 1. The OM content for the

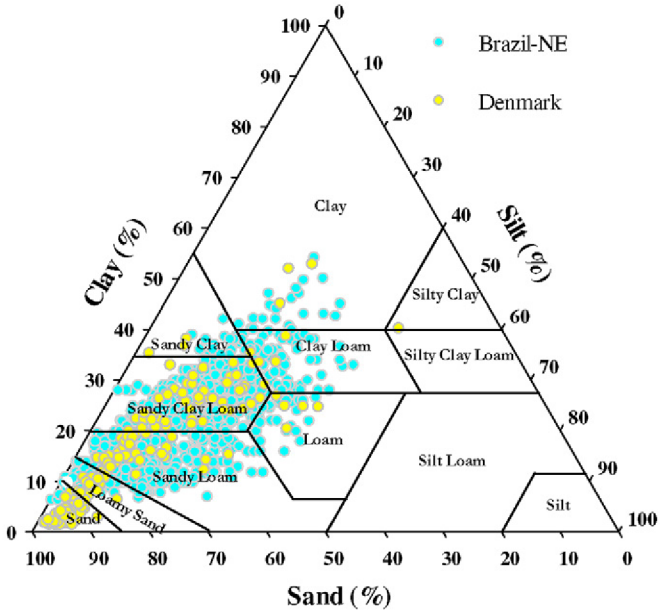

Fig. 2. Soils in the Brazil-NE and Denmark datasets used to develop a pedotransfer function PTF-θ(*h*) for the prediction of water content at specific pressure heads on a USDA textural triangle.

Danish sites is greater than for the Brazilian soils due to climatic conditions (temperate subhumid vs. tropical semiarid) as well as their intensive farming background. Bulk density is just slightly variable across both datasets, with a coefficient of variation (CV) of 10% on average. The spatial variability of OM, water content at the wilting point (θ_{150}), and clay content are larger for the Danish

Table 1. Descriptive analysis of soil samples from the Denmark and Brazil-NE datasets.

Soil property	Dataset	Min.	Max.	Mean	Median	CV	Kurtosis	Skewness
						%		
Organic matter, kg kg^{-1}	Denmark	0	0.100	0.015	0.008	160	8.02	1.87
	Brazil-NE	0	0.027	0.006	0.005	67	6.57	1.66
Sand, kg kg^{-1}	Denmark	0.18	0.97	0.74	0.78	24	2.61	−0.68
	Brazil-NE	0.25	0.96	0.64	0.64	22	2.50	−0.26
Clay, kg kg^{-1}	Denmark	0.01	0.42	0.10	0.07	80	4.80	1.42
	Brazil-NE	0.02	0.54	0.22	0.22	41	2.67	0.40
Bulk density, Mg m^{-3}	Denmark	1.18	1.99	1.54	1.54	10	2.77	0.23
	Brazil-NE	0.73	1.98	1.69	1.70	8	6.50	−0.12
$\theta_{0.1}$, m^3 m^{-3}	Denmark	0.24	0.51	0.38	0.38	13	2.85	0.07
$\theta_{0.6}$, m^3 m^{-3}	Brazil-NE	0.10	0.69	0.27	0.27	33	5.09	0.85
θ_{1}, m^3 m^{-3}	Denmark	0.02	0.43	0.22	0.23	45	2.14	−0.01
	Brazil-NE	0.036	0.59	0.21	0.21	38	4.73	0.83
θ_{3}, m^3 m^{-3}	Brazil-NE	0.014	0.52	0.17	0.17	47	4.94	0.91
θ_{5}, m^3 m^{-3}	Brazil-NE	0.013	0.50	0.16	0.16	44	5.17	0.96
θ_{10}, m^3 m^{-3}	Denmark	0.01	0.37	0.14	0.13	57	2.33	0.46
	Brazil-NE	0.012	0.47	0.15	0.15	47	4.82	0.88
θ_{20}, m^3 m^{-3}	Brazil-NE	0.012	0.46	0.14	0.14	50	4.70	0.85
θ_{150}, m^3 m^{-3}	Denmark	0.00	0.31	0.06	0.04	83	7.29	1.58
	Brazil-NE	0.011	0.39	0.11	0.11	45	4.32	0.76

soils, with respective CVs of 160, 83, and 80% compared with 67, 45, and 41% for the Brazilian samples. The skewness is low, except for some properties (OM, clay content, and θ_{150}). Based on the sensitivity of skewness to extreme values, discussed also by Isaaks and Srivastava (1989), we conclude that there is no local distribution of soil properties. In general, the median and mean values are almost the same, suggesting a normal distribution of the data (Nielsen and Wendroth, 2003). In both datasets, OM and clay contents are positively skewed while sand content is negatively skewed.

Water contents are typically correlated with texture, OM, and BD, and this has been the basis for using regression analysis to develop PTFs. Table 2 shows the correlation of the measured water contents that are common in both datasets (Brazil-NE and Denmark). Significant differences between correlations can be observed, for example, in the near-saturated water content (θ_{min}, considered as $\theta_{0.1}$ for Denmark and as $\theta_{0.6}$ for Brazil-NE). Danish soils showed θ_{min} to be positively correlated to OM ($r = 0.57$) and negatively to BD ($r = -0.48$). However, this is not the case for soils from Brazil where θ_{min} is significantly correlated only with particle size fractions ($r = -0.86$ with sand content and 0.83 with clay content). For Danish soils, the correlation between water contents at higher pressure heads and texture is stronger. A similar observation can be made for the soils in Brazil-NE. Clay particles are associated with small pore diameters and a large surface area, leading to an increase in water adsorption (Martin, 1962). Clay is therefore a very good predictor for water contents in the drier part of the soil moisture retention range. A positive correlation between OM and drier water contents can be observed for the Danish soils.

Covariate Shift Detection

Covariate shift, one of the common biases in data when training and testing datasets present different distributions (different means and variances), as assessed by shuffling the data and randomly selecting the training set, resulted in 100 PTFs. The mean and variance of predictors (OM, sand content, clay content, and BD) for the 100 random repetitions are shown in Fig. 3 for the Danish soils. Scattering in mean and variance values suggests the existence of covariate shift in predictors during the PTF development. This procedure was implemented for the dataset Brazil-NE as well.

To test for covariate shift, the two-sample Kolmogorov–Smirnov test was applied. This test assesses the equality of two probability distribution functions considering the null hypothesis (H_0), which confirms that the distributions originate from the same probability distribution function, contrary to the alternative hypothesis (H_1). Furthermore, Kolmogorov–Smirnov describes the difference between either the mean value or the variance of the two selected samples (Cieslak and Chawla, 2009). Therefore, covariate shift in the developed PTFs for both regions can be shown to exist by assessing the probability of the alternative hypothesis in all 100 shuffled datasets in PTF predictors OM, sand, clay, and BD. Note that the existence of covariate shift in at least one of the predictors is enough to result in statistical bias in PTF development.

Table 3 shows the mean value and standard deviation of the probability of the alternative hypothesis (i.e., covariate shift) obtained by Kolmogorov–Smirnov for each predictor used for the development of PTFs in both datasets. Each predictor showed a probability in the order of 30 to 40% of experiencing covariate shift by data shuffling.

Figure 4 shows the distribution of the probability of the occurrence of covariate shift in n predictors, with n varying between 0 (no predictor with covariate shift) and 4 (all predictors—OM, sand, clay and BD—with covariate shift) during development of PTFs for the Denmark and Brazil-NE datasets. In the vast majority of cases, covariate shift occurred in at least one of the predictors, and only about 7 ± 2 (Denmark) and 10 ± 4.5% (Brazil-NE) of samples had no covariate shift according to the Kolmogorov–Smirnov test. About 70% of all sample cases showed covariate shift in one or two predictors.

Pedotransfer Function for Water Retention

The random selection of the training datasets induced different distributions in each of the 100 shuffling repetitions. As an example, Fig. 5 shows the uncertainty in the prediction of $\theta_{0.1}$ and θ_{150} for the Danish soils, representing wet and dry conditions, using the four development methods (LM, SLM, GP, and ENS).

The performance of all developed PTFs for the prediction of $\theta_{0.1}$ and θ_{150} in terms of RMSE varied significantly due to variation in the training and testing sample distributions (Fig. 5). For near-saturated water content ($\theta_{0.1}$), the ENS-PTF was the best PTF, as the RMSE varied between 0.030 and 0.055 $m^3\ m^{-3}$ with

Table 2. Pearson correlation analysis between water retention at the pressure head near saturation (θ_{min}), and at pressure heads of −1 m (θ_1), −10 m (θ_{10}), and −150 m (θ_{150}) with organic matter (OM), particle size distribution (clay and sand contents), and bulk density (BD) for the Brazil-NE and Denmark datasets.

	θ_{min}†		θ_1		θ_{10}		θ_{150}	
Soil property	Brazil-NE	Denmark	Brazil-NE	Denmark	Brazil-NE	Denmark	Brazil-NE	Denmark
OM	−0.03	0.57	0.01	0.49	0.01	0.40	0.01	0.29
Sand	−0.86	0.04	−0.86	−0.76	−0.87	−0.81	−0.87	−0.81
Clay	0.83	−0.15	0.86	0.63	0.88	0.77	0.89	0.85
BD	0.07	−0.48	0.01	−0.20	−0.02	−0.14	−0.02	−0.10

† θ_{min} is the water content at the minimum available tension for each dataset: $\theta_{0.6}$ for Brazil-NE and $\theta_{0.1}$ for Denmark.

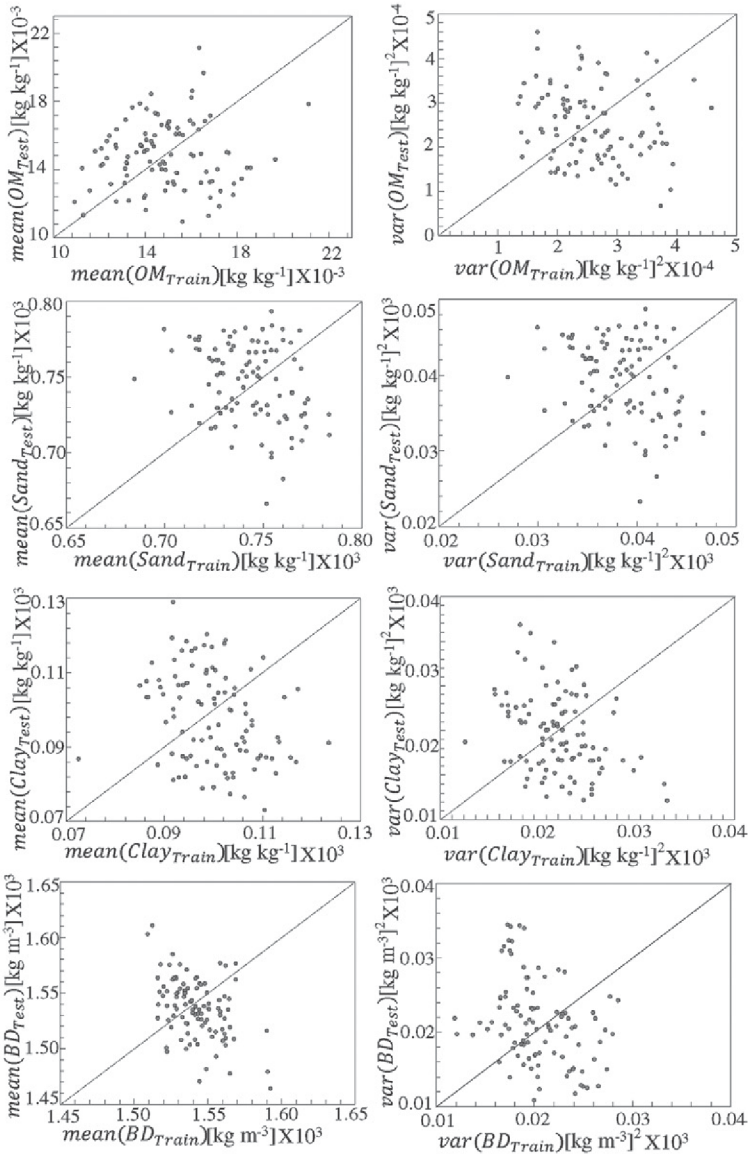

Fig. 3. Scatterplots of the mean and variance of predictors (organic matter [OM], sand content, clay content, and bulk density [BD]) used in training and testing of 100 data shuffles for soils from Denmark.

an average value of 0.041 $m^3\ m^{-3}$, followed by the GPR-PTF and SLM-PTF, with equal average RMSE values of 0.044 $m^3\ m^{-3}$. This robust performance of the developed ensemble PTF with a low number of predictors is as good as the PTFs presented by Børgesen and Schaap (2005), who used seven classes of Danish soil texture. However, in case of the ENS-PTF, the performance decreased strongly for the testing data, probably indicating overtraining.

Table 3. Kolmogorov–Smirnov statistical test (p value of 0.05) on the training and testing datasets for 100 data shuffles for both regions (Denmark and Brazil-NE).

Variable	Probability of covariate shift in predictor	
	Denmark	Brazil-NE
	%	
Organic matter	29.6 ± 2.4	33.2 ± 4.4
Sand	37.8 ± 7.8	44.2 ± 5.5
Clay	33.8 ± 4.5	38.2 ± 3.8
Bulk density	29.6 ± 2.4	30.4 ± 4.3

For the wilting point (θ_{150}), model complexity did not result in better predictions because LM and SLM not only had lower RMSE values (0.020 $m^3\ m^{-3}$) compared with the GP and ENS methods but also less uncertainty. These results are comparable with errors obtained from European parametric PTFs proposed by Vereecken et al. (1989) and Tóth et al. (2015) for the prediction of water content at the wilting point, although they included cation exchange capacity among the predictors. The LM-PTF for the prediction of θ_{150} with the uncertainty obtained for Danish soils confirms the high importance of clay content and OM, as also observed by Wösten et al. (2001) and Rawls et al. (2003). It is

$$\theta_{150} = (0.65 \pm 0.17)\mathrm{OM} - (0.01 \pm 0.02)\mathrm{Sand} + (0.47 \pm 0.06)\mathrm{Clay} + (0.00 \pm 0.02)\mathrm{BD} \quad [7]$$

The observed variability in performance is rooted in the variability of the model compartments, i.e., the coefficients of predictors in the LM-PTF and SLM-PTF, such as the means and standard deviations shown in Eq. [7], and the relative importance of the predictors in the GP-PTF and ENS-PTF. Figure 6a shows the variation in the relative importance of predictors in the GP-PTF and ENS-PTF resulting from the 100 simulations. The distribution of the importance of predictors in the GP-PTF is larger than in the ENS-PTF for prediction of $\theta_{0.1}$, showing that GP regression is very sensitive to covariate shift. In GP regression (Fig. 6a), OM played a more important role in the prediction of $\theta_{0.1}$ (relative importance of 0.30) than clay content and BD (relative importance of 0.25), whereas sand content was least important. There is also a substantial standard deviation in the relative importance of BD and sand content, confirming that the role of these parameters in prediction is sensitive to covariate shift.

The ensemble method (ENS) used mainly OM as a predictor (Fig. 6b), with a relative importance of 0.47 ± 0.15, while the remainder was equally divided among the other three predictors. The predictor importance for the wilting point was not shown because its best PTF (LM) was already presented (Eq. [7]).

For the Brazil-NE dataset, Fig. 7 shows the results for predicted water contents $\theta_{0.6}$ and θ_{150}. Due to the strong linear correlation between texture and water contents (Table 2), both the LM-PTF and SLM-PTF were able to predict $\theta_{0.6}$ precisely, with RMSE values of 0.040 and 0.036 $m^3\ m^{-3}$, respectively. Machine learning methods including GP and ENS showed the same performance as LM; however, ENS was overtrained, as shown by the disparity between training and testing RMSE values. The PTFs developed by Tomasella et al. (2003) for Brazilian soils predicted

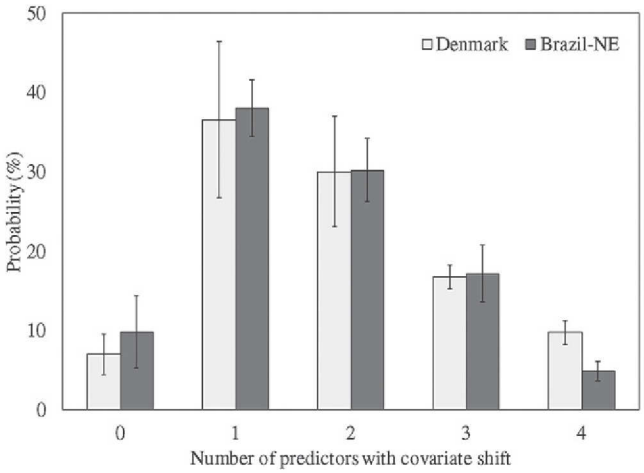

Fig. 4. Probability of the number of predictors (organic matter, sand, clay, and bulk density) with covariate shift during development of pedotransfer functions for the Denmark and Brazil-NE datasets.

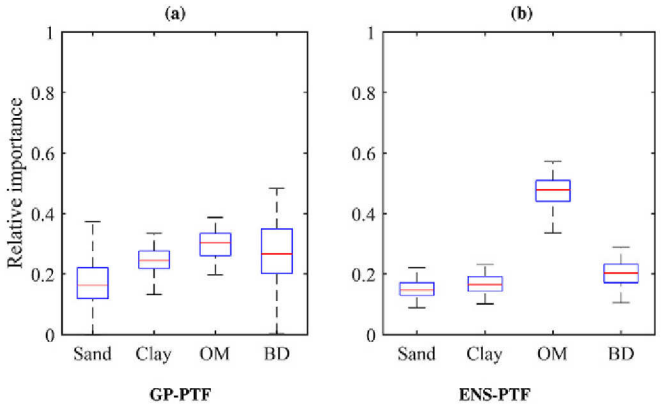

Fig. 6. Box-and-whisker plots showing the relative importance of predictors (organic matter [OM], sand content, clay content, and bulk density [BD]) in the pedotransfer functions developed by (a) Gaussian process regression (GP-PTF) and (b) ensemble with least square boosting (ENS-PTF) for the prediction of water content at −0.1 m pressure head ($\theta_{0.1}$) in soils from Denmark obtained from 100 simulations with random selection of the training datasets.

$\theta_{0.6}$ with an RMSE of 0.046 m^3 m^{-3} but require detailed information about the particle size distribution (fine and coarse sand fractions), which is not always available. The RMSE values of 0.046 and 0.050 m^3 m^{-3} for the prediction of this water content were also obtained by Barros et al. (2013) using soil samples from the northeast of Brazil. For sandy and clayey soils from different parts of Brazil, da Silva et al. (2017) reported RMSE values of

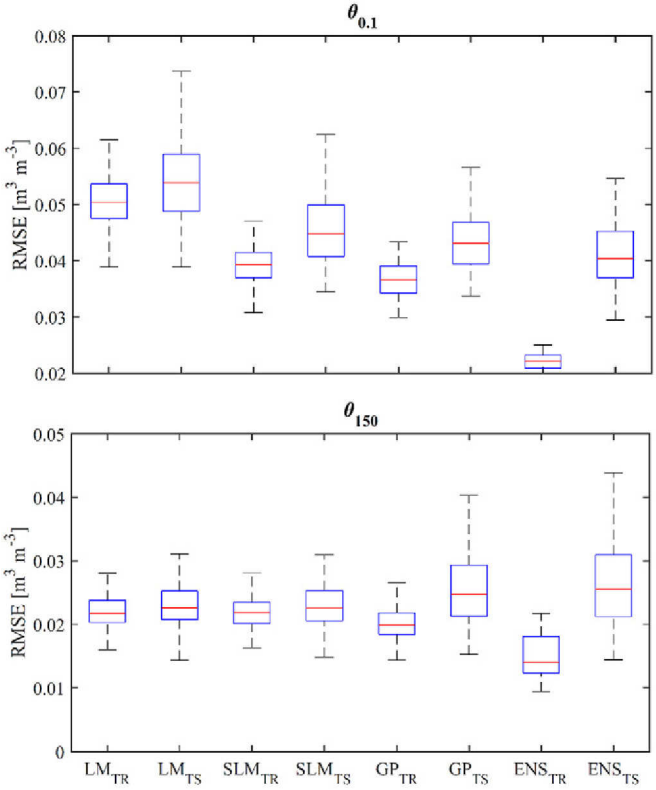

Fig. 5. Box-and-whisker plots of RMSE values for water content at −0.1 and −150 m pressure heads ($\theta_{0.1}$ and θ_{150}, respectively) obtained from comparison of estimated values from the 100 developed pedotransfer functions with the observed values for soils from Denmark using four development methods: linear regression (LM), stepwise linear regression (SLM), Gaussian process regression (GP), and ensemble with least square boosting (ENS); subscripts TR and TS refer to the training and testing datasets.

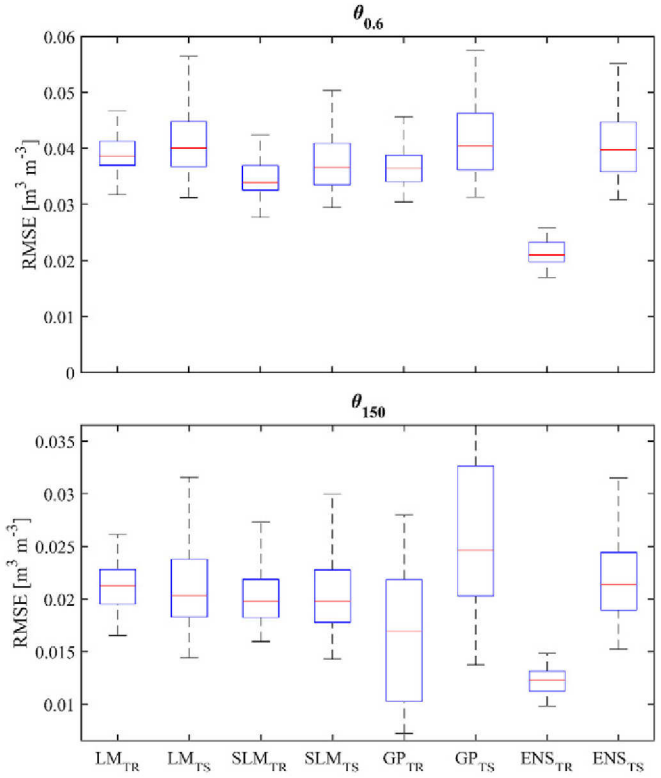

Fig. 7. Box-and-whisker plots of RMSE values for water content at −0.6 and −150 m pressure heads ($\theta_{0.6}$ and θ_{150}, respectively) obtained from comparison of estimated values from the 100 developed pedotransfer functions for soils from Brazil-NE using four development methods: linear regression (LM), stepwise linear regression (SLM), Gaussian process regression (GP) and ensemble with least square boosting (ENS); subscripts TR and TS refer to the training and testing datasets.

about 0.030 and 0.050 $m^3\ m^{-3}$ obtained by predictions from a semi-deterministic PTF.

The GP-PTF was not able to predict the wilting point under covariate shift with the same accuracy as the other methods. The obtained RMSE of 0.020 $m^3\ m^{-3}$ is similar to the results of van den Berg et al. (1997) and Barros et al. (2013), who developed PTFs for Brazilian soils. The important feature of regression-based methods is the similarity of performance prediction between training and testing stages, a feature not observed in ENS and explained by overtraining. Overtraining is revealed by the very low values for RMSE using this algorithm—0.012 $m^3\ m^{-3}$ on average, while the RMSE increased to 0.022 $m^3\ m^{-3}$ in the testing phase.

Robust and simple LMs predicting $\theta_{0.6}$ and θ_{150} for the Brazil-NE dataset were

$$\theta_{0.6} = (2.32 \pm 0.57)\mathrm{OM} - (0.27 \pm 0.03)\mathrm{Sand} + (0.47 \pm 0.04)\mathrm{Clay} + (0.19 \pm 0.08)\mathrm{BD} \quad [8]$$

$$\theta_{150} = (1.66 \pm 0.22)\mathrm{OM} - (0.17 \pm 0.02)\mathrm{Sand} + (0.38 \pm 0.02)\mathrm{Clay} + (0.08 \pm 0.01)\mathrm{BD} \quad [9]$$

The greater effect of BD for predicting the water content near saturation, as observed when comparing these equations, was already confirmed by Rawls et al. (2003) and Nguyen et al. (2017).

Figure 8a shows a very large variability in the predictors of the GP-PTF for $\theta_{0.6}$ and suggests that the effect of OM in prediction of this water content can be ignored. However, removal of this variable increased the RMSE for the prediction of $\theta_{0.6}$ to 0.044 $m^3\ m^{-3}$ in the testing phase. Sand, clay, and BD shared the relative importance in prediction of $\theta_{0.6}$ at 0.46, 0.28, and 0.20 on average, respectively. However, this large variability did not occur in components of the ENS-PTF (Fig. 8b) when they were exposed to covariate shift, and texture-based predictors were mainly used to predict $\theta_{0.6}$, specifying about 90% of the relative importance together. In view of these findings, $\theta_{0.6}$ was also predicted using the ENS-PTF method with only sand and clay as predictors. Compared with the models using all the predictors, the RMSE increased by 24%. Therefore, in this case, the elimination of predictors is not recommended. We did not show these results for wilting point prediction where linear models were much more accurate than GP and ENS.

The best PTF was identified by comparing R^2 and RMSE values for each water content in the Denmark and Brazil-NE datasets (Table 4). For the Danish dataset, ENS-PTF performed better than other algorithms for the prediction of $\theta_{0.1}$ and θ_1. However, the prediction of $\theta_{0.1}$ was unreliable, with low and largely unstable values of R^2 ranging between 0.18 to 0.60. Due to the high relative importance of clay content in the prediction of θ_{150} for the Brazilian dataset, LM was the most accurate model, with an RMSE of 0.020 $m^3\ m^{-3}$ and a small interval of variation of R^2.

The SLM-PTF was the best approach with the highest R^2 (0.85 on average) for the prediction of water contents for the Brazil-NE soils. Based on these findings, it can be interpreted that simple regression-based PTFs perform better than complex machine learning models when the intercorrelation of water contents is high. These PTFs require less computational effort for simulation.

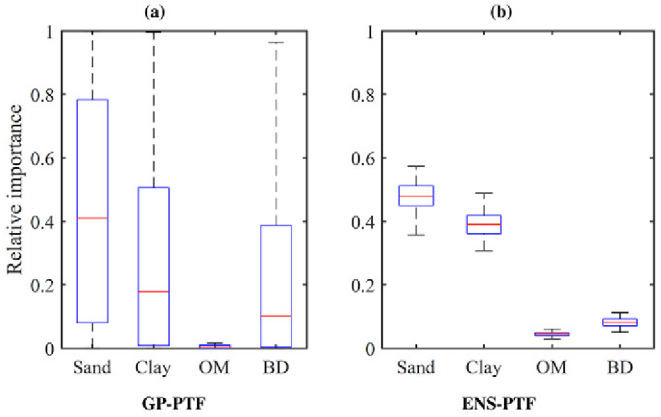

Fig. 8. Box-and-whisker plots showing the relative importance of predictors (organic matter [OM], sand content, clay content, and bulk density [BD]) in the pedotransfer functions developed by (a) Gaussian process regression (GP-PTF) and (b) ensemble with least square boosting (ENS-PTF) for the prediction of water content at −0.6 m pressure head ($\theta_{0.6}$) in soils from northeastern Brazil obtained from 100 simulations with random selection of the training datasets.

Figure 9 shows the 100 PTF-generated soil water retention curves for four soils of different texture classes from the Danish dataset, together with the curve fitted to the measured data and the associated van Genuchten (1980) parameters' probability distribution functions. The underestimation of θ_{150} in the clay loam and sandy clay loam (Fig. 9a and 9b), together with the large variability due to the random selection of the training dataset yielded discrepancies in the van Genuchten parameters (Table 5).

Table 4. Best pedotransfer function (PTF) selected for the prediction of water contents at specific pressure heads in Denmark and Brazil (NE): ensemble regression with bagging aggregation PTF (ENS), stepwise linear PTF (SLM), Gaussian regression PTF (GP), and linear PTF (LM).

Target†	Dataset	Best PTF	R^2	RMSE
$m^3\ m^{-3}$				$m^3\ m^{-3}$
$\theta_{0.1}$	Denmark	ENS	0.35 (0.18–0.60)‡	0.041 (0.030–0.051)
$\theta_{0.6}$	Brazil-NE	SLM	0.82 (0.68–0.88)	0.036 (0.027–0.054)
θ_1	Denmark	ENS	0.73 (0.45–0.84)	0.051 (0.034–0.073)
	Brazil-NE	SLM	0.85 (0.71–0.92)	0.031 (0.021–0.046)
θ_3	Brazil-NE	SLM	0.88 (0.70–0.93)	0.026 (0.018–0.046)
θ_5	Brazil-NE	SLM	0.87 (0.76–0.92)	0.026 (0.013–0.046)
θ_{10}	Denmark	GP	0.73 (0.13–0.93)	0.041 (0.023–0.086)
	Brazil-NE	SLM	0.86 (0.71–0.95)	0.026 (0.019–0.037)
θ_{20}	Brazil-NE	SLM	0.86 (0.72–0.95)	0.024 (0.019–0.037)
θ_{150}	Denmark	SLM	0.72 (0.17–0.91)	0.021 (0.016–0.043)
	Brazil-NE	LM	0.87 (0.71–0.92)	0.020 (0.015–0.032)

† $\theta_{0.1}$, $\theta_{0.6}$, θ_1, θ_3, θ_5, θ_{10}, θ_{20}, and θ_{150} are water contents at pressure heads of −0.1, −0.6, −1, −3, −5, −10, −20, and −150 m, respectively.
‡ Range (minimum–maximum) in parentheses.

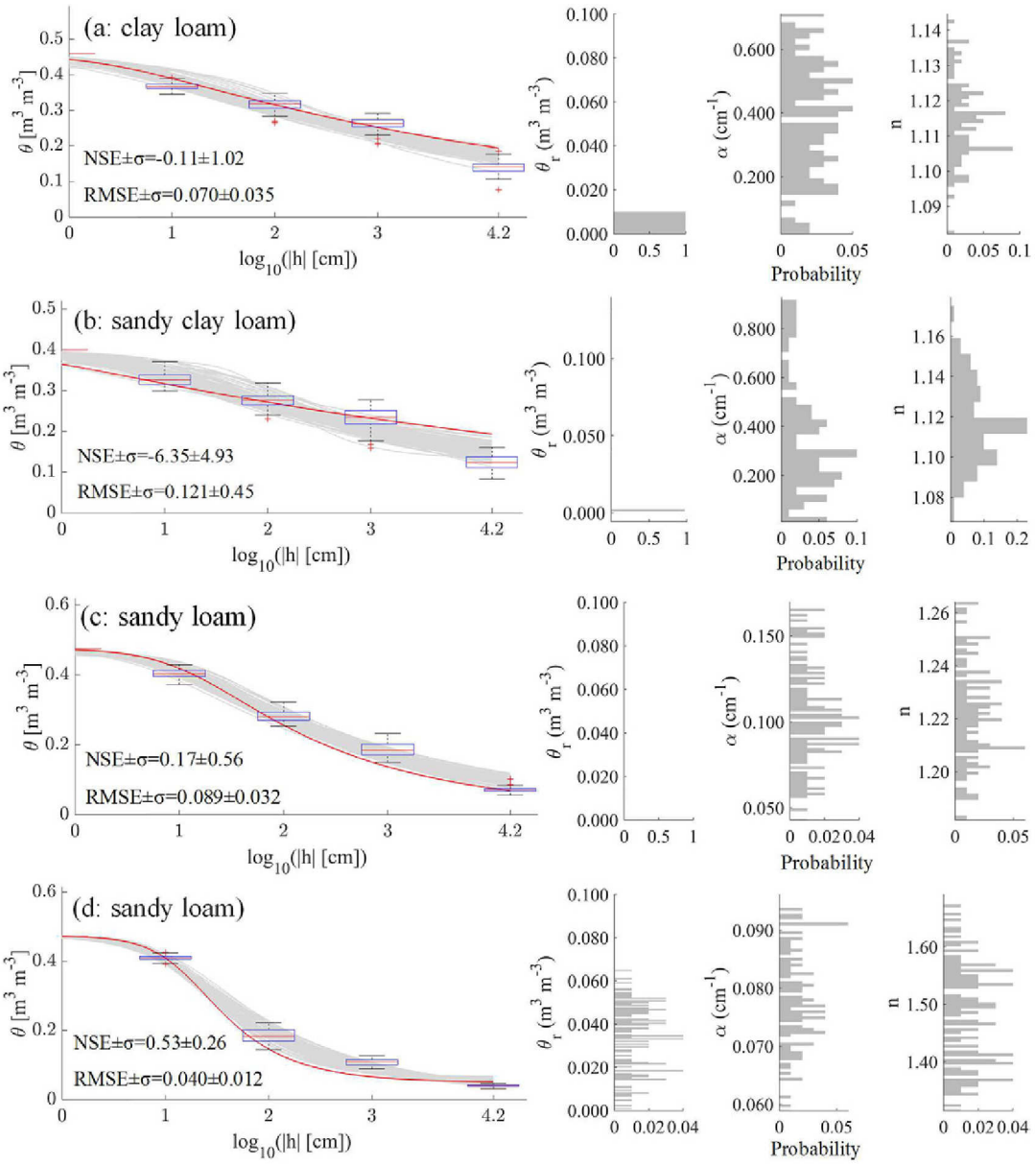

Fig. 9. Soil water retention curves for (a) a clay loam, (b) a sandy clay loam, (c) a sandy loam, and (d) a sandy soil from the Danish dataset. Gray lines are plotted with 100 simulations from the estimated points with the best pedotransfer functions (PTFs) shown in Table 4, while the red line was fitted from measured data. The box plots are related to points predicted from the PTFs. The histograms are related to the respective van Genuchten parameters residual water content (θ_r), α, and n from the simulations.

Table 5. Average van Genuchten parameters† for soils from the Danish dataset for the simulated scenario using the results of 100 soil water retention curves fitted to the water contents generated with the best PTFs from Table 4 and for the observed scenario with parameters obtained from measured data.

Soil sample	Soil type	Scenario	θ_r	θ_s	α	n
			$m^3\ m^{-3}$		cm^{-1}	
a	clay loam	simulated	0.000	0.456	0.408 (0.232)‡	1.117 (0.017)
		observed	0.000	0.456	0.419	1.098
b	sandy clay loam	simulated	0.003	0.39	0.421 (0.455)	1.127 (0.078)
		observed	0.00	0.40	3.054	1.068
c	sandy loam	simulated	0.001 (0.002)	0.471	0.107 (0.034)	1.224 (0.023)
		observed	0.014	0.476	0.0819	1.300
d	sandy	simulated	0.034 (0.015)	0.474	0.080 (0.010)	1.494 (0.110)
		observed	0.050	0.474	0.064	1.787

† θ_r, residual water content; θ_s, saturated water content; α, scale parameter; n, shape parameter.
‡ Standard deviations in parentheses.

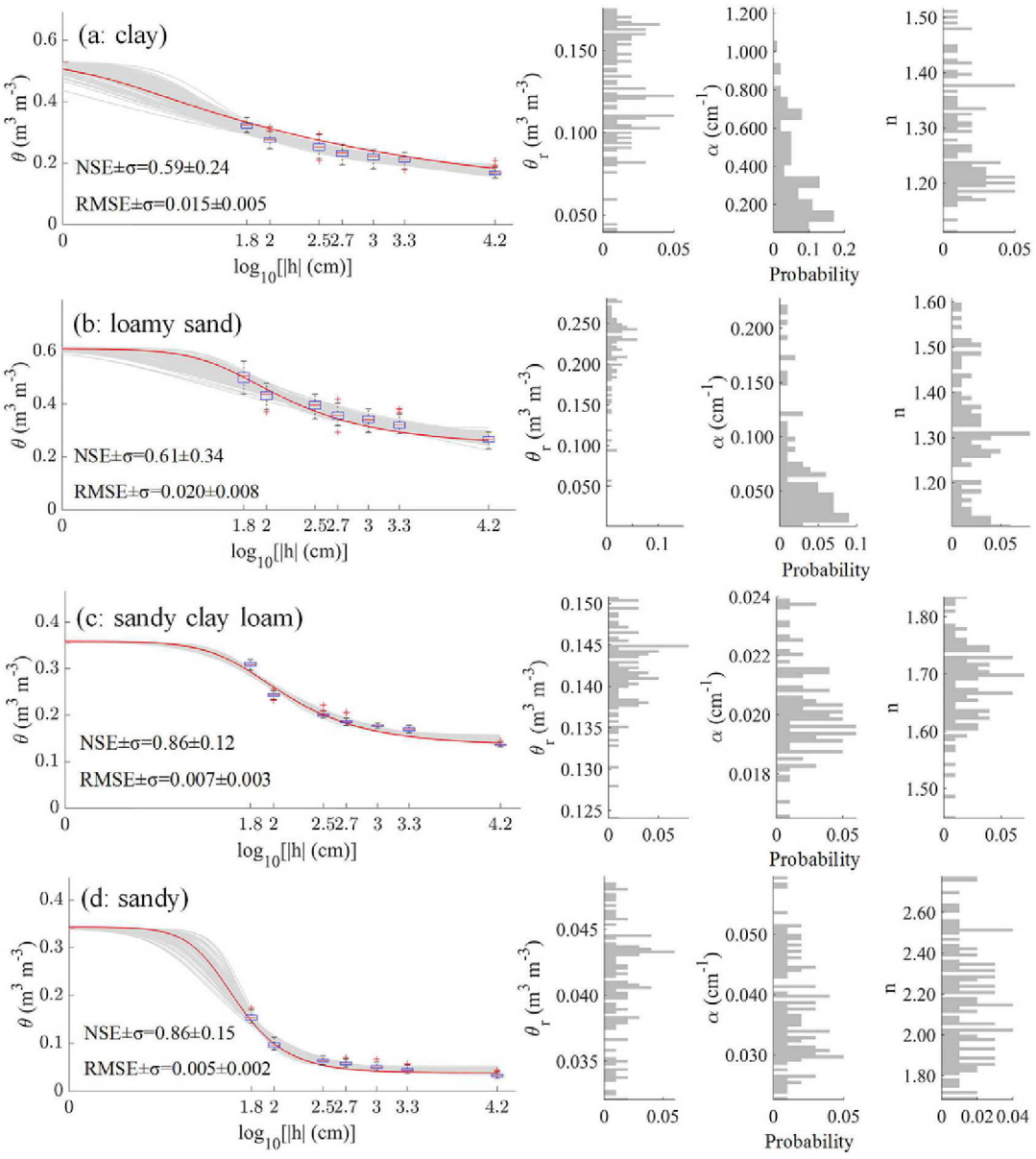

Fig. 10. Generated soil water retention curves for (a) a clay, (b) a loamy sand, (c) a sandy clay loam, and (d) a sandy soil from the Brazil-NE dataset. Gray lines represent 100 fits obtained from the best pedotransfer functions (PTFs) shown in Table 4, while the red line was fitted to measured data. The box plots refer to water contents predicted from the PTFs. The histograms are related to the respective van Genuchten parameters residual water content (θ_r), α, and *n* from the PTF predictions.

For the clay loam, although the mean α value obtained from the PTF and the observed data were similar (0.408 ± 0.232 and 0.419 cm^{-1}), the standard deviations in scenarios from estimated points were high (Table 5). Additionally, there is no specific distribution for α values (Fig. 9a and 9b).

There are two issues when using PTFs to estimate water content in sandy clay loam soils (Fig. 9b). First, all PTFs yielded about 10% underestimation for θ_{150}. Second, although the observed water contents, especially θ_1 and θ_{10}, crossed the center of boxes for simulated data (Fig. 9b), longer whiskers for each predicted water content were observed. Thus, very different α values (0.421 ± 0.455 cm^{-1}) were fitted using PTF predictions compared with the observed data (α = 3.05 cm^{-1}) also shown in Table 5. There were large RMSE values and negative values of NSE for both soils (Fig. 9a and 9b), where the predicted water retention using PTF points explains the deficiency of the PTF-generated points for proper estimation of the curve.

In sandy loam and sandy soils (Fig. 9c and 9d), the more accurate estimation of $\theta_{0.1}$, θ_1, and θ_{10} positively affected the estimation of α (Table 5), but in both cases θ_{150} was underestimated. A better estimation of water content for soils with a higher sand content was expected because of the larger number of sandy soil samples in the dataset, yielding better trained PTFs.

In the same way as for the Danish dataset, the uncertainty of the developed PTFs for the Brazil-NE dataset was evaluated.

Table 6. Average van Genuchten parameters† for soils from Brazil-NE dataset for the simulated scenario (results of 100 soil water retention curves fitted to the water contents generated with the best PTFs from Table 4) and for the observed scenario using parameters obtained from measured data.

Soil sample	Soil type	Scenario	θ_r	θ_s	α	n
			$m^3\ m^{-3}$		cm^{-1}	
a	clay	simulated	0.185 (0.091)‡	0.611	0.077 (0.079)	1.322 (0.173)
		observed	0.245	0.610	0.0248	1.515
b	loamy sand	simulated	0.128 (0.033)	0.532 (0.028)	0.460 (0.661)	1.293 (0.117)
		observed	0.081	0.532	0.551	1.166
c	sandy clay loam	simulated	0.142 (0.006)	0.357	0.021 (0.002)	1.694 (0.089)
		observed	0.135	0.359	0.0203	1.667
d	sandy	simulated	0.041 (0.005)	0.340	0.039 (0.011)	2.208 (0.328)
		observed	0.038	0.343	0.036	2.242

† θ_r, residual water content; θ_s, saturated water content; α, scale parameter; n, shape parameter.
‡ Standard deviations in parentheses.

One hundred retention curves based on PTF predictions of the respective water contents were generated for four randomly selected soils in the classes clay, loamy sand, sandy clay loam, and sandy. The retention curves and respective histograms of the van Genuchten parameters are presented in Fig. 10, showing less variation in the estimated water contents than in the Danish dataset. This result confirms the average low RMSE values of water contents predicted by the best PTFs for the Brazilian dataset: <3% (Table 4).

Due to the poor performance of the PTFs for clay and loamy sand soils (Fig. 10a and 10b), not only was the scatter in the fitted van Genuchten parameters higher, but there was also an increase in RMSE and a decrease in NSE values. For the sandy and sandy clay loam (Fig. 10c and 10d), NSE was >0.8. The highest obtained α was about 0.3 cm^{-1} for the clay soil and 0.1 cm^{-1} for the loam soil, while the true values from experimental data were 0.08 and 0.55 cm^{-1}, respectively (Table 6). In the clay and loamy sand soils, fitted values of α were in the range of 0.35 and 0.06 cm^{-1}, respectively.

Whiskers for the predicted water contents are short, and in each texture class, some water contents were predicted with high accuracy (Fig. 10c and 10d). This lower variability and accurate prediction was reflected in similar retention parameters using a PTF or observed data. The large value of NSE (0.8) and RMSE values close to zero confirm this, as well as very similar values of the van Genuchten parameters as the average of the PTF-predicted curves and the observed curve (Table 6). For most cases, both α and n values are very similar for simulated and observed scenarios.

Conclusions

In this study, we investigated the effect of covariate shift due to random shuffling of the data before PTF development on the performance of the PTF. Linear and stepwise regressions and machine learning methods including Gaussian process and ensemble regression were used to develop PTFs. The methodology was applied to water contents at corresponding pressure heads using datasets of soils from Denmark and Brazil. The conclusions of this study are:

1. Shuffling of data leading to covariate shift results in uncertainty in the prediction of water contents by the developed PTFs.
2. Inherent variability of data as observed in the Danish dataset may lead to increased prediction uncertainty.
3. For correlated data such as water contents from the Brazil-NE dataset, simple regression models performed as well as sophisticated machine learning methods.
4. Using PTF-predicted water contents for water retention curve fitting may lead to high uncertainty in the van Genuchten parameters.

Data Availability

Data from this study are available through the Dryad Data Repository (Kotlar et al., 2019a).

Acknowledgments

Ali Mehmandoost Kotlar is grateful to the FAPESP Foundation, Brazil, for providing financial support (scholarship) for his study (process no. 2016/18636-7). The support from the Danish Environment Agency financed project Mapping of Risk Areas for Phosphorus Loss and Phosphorus Sensitive Water Areas (Fosforkortlaegning) and from the Danish Pesticide Leaching Assessment Programme (pesticidvarsling.dk) is gratefully acknowledged.

References

Baker, L., and D. Ellison. 2008. Optimisation of pedotransfer functions using an artificial neural network ensemble method. Geoderma 144:212–224. doi:10.1016/j.geoderma.2007.11.016

Barros, A.H.C., Q. de Jong van Lier, A.H.N. Maia, and F.V. Scarpare. 2013. Pedotransfer functions to estimate water retention parameters of soils in northeastern Brazil. Rev. Bras. Cienc. Solo 37:379–391. doi:10.1590/S0100-06832013000200009

Bishop, C.M. 2006. Pattern recognition and machine learning. Springer, New York.

Børgesen, C.D., B.V. Iversen, O.H. Jacobsen, and M.G. Schaap. 2008. Pedotransfer functions estimating soil hydraulic properties using different soil parameters. Hydrol. Processes 22:1630–1639. doi:10.1002/hyp.6731

Børgesen, C.D., and M.G. Schaap. 2005. Point and parameter pedotransfer functions for water retention predictions for Danish soils. Geoderma 127:154–167. doi:10.1016/j.geoderma.2004.11.025

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. Adv. Soil Sci. 9:177–213. doi:10.1007/978-1-4612-3532-3_4

Campos de Oliveira, M.H., V. Sari, N.M. de Reis Castro, and O.C. Pedrollo. 2017. Estimation of soil water content in watershed using artificial neural networks. Hydrol. Sci. J. 62:2120–2138. doi:10.1080/02626667.2017.1364844

Chirico, G.B., H. Medina, and N. Romano. 2010. Functional evaluation of PTF prediction uncertainty: An application at hillslope scale. Geoderma 155:193–202. doi:10.1016/j.geoderma.2009.06.008

Chung, Y., P.J. Haas, E. Upfal, and T. Kraska. 2018. Unknown examples & machine learning model generalization. arXiv1808.08294 [cs.LG].

Cichota, R., I. Vogeler, V.O. Snow, and T.H. Webb. 2013. Ensemble pedotransfer functions to derive hydraulic properties for New Zealand soils. Soil Res. 51:94–111. doi:10.1071/SR12338

Cieslak, D.A., and N.V. Chawla. 2009. A framework for monitoring classifiers' performance: When and why failure occurs? Knowl. Inf. Syst. 18:83–108. doi:10.1007/s10115-008-0139-1

da Silva, A.C., R.A. Armindo, A. dos Santos Brito, and M.G. Schaap. 2017. SPLINTEX: A physically-based pedotransfer function for modeling soil hydraulic functions. Soil Tillage Res. 174:261–272. doi:10.1016/j.still.2017.07.011

D'Emilio, A., R. Aiello, S. Consoli, D. Vanella, and M. Iovino. 2018. Artificial neural networks for predicting the water retention curve of Sicilian agricultural soils. Water 10:1431. doi:10.3390/w10101431

Deng, H., M. Ye, M.G. Schaap, and R. Khaleel. 2009. Quantification of uncertainty in pedotransfer function-based parameter estimation for unsaturated flow modeling. Water Resour. Res. 45:W04409. doi:10.1029/2008WR007477

Efron, B., and R.J. Tibshirani. 1994. An introduction to the bootstrap. CRC Press, Boca Raton, FL.

Guber, A.K., Y.A. Pachepsky, M.Th. van Genuchten, W.J. Rawls, J. Šimůnek, D. Jacques, et al. 2006. Field-scale water flow simulations using ensembles of pedotransfer functions for soil water retention. Vadose Zone J. 5:234–247. doi:10.2136/vzj2005.0111

Gupta, S.C., and W.E. Larson. 1979. Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. Water Resour. Res. 15:1633–1635. doi:10.1029/WR015i006p01633

Isaaks, E.H., and R.M. Srivastava. 1989. An introduction to applied geostatistics. Oxford Univ. Press, New York.

Iversen, B.V., C.D. Børgesen, M. Lægdsmand, M.H. Greve, G. Heckrath, and C. Kjærgaard. 2011. Risk predicting of macropore flow using pedotransfer functions, textural maps, and modeling. Vadose Zone J. 10:1185–1195. doi:10.2136/vzj2010.0140

Jarvis, N., J. Koestel, I. Messing, J. Moeys, and A. Lindahl. 2013. Influence of soil, land use and climatic factors on the hydraulic conductivity of soil. Hydrol. Earth Syst. Sci. 17:5185–5195. doi:10.5194/hess-17-5185-2013

Khlosi, M., M. Alhamdoosh, A. Douaik, D. Gabriels, and W.M. Cornelis. 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. Eur. J. Soil Sci. 67:276–284. doi:10.1111/ejss.12345

Kotlar, A.M., Q. de Jong van Lier, A.H.C. Barros, B.V. Iversen, and H. Vereecken. 2019a. Dataset for: Development and uncertainty assessment of pedotransfer functions for predicting water contents at specific pressure heads. Dryad Data Repository. doi:10.5061/dryad.3r2280gbw

Kotlar, A.M., B.V. Iversen, and Q. de Jong van Lier. 2019b. Evaluation of parametric and nonparametric machine-learning techniques for prediction of saturated and near-saturated hydraulic conductivity. Vadose Zone J. 18:180141. doi:10.2136/vzj2018.07.0141

Kotlar, A.M., I. Varvaris, Q. de Jong van Lier, L.W. de Jonge, P. Moldrup, and B.V. Iversen. 2019c. Soil hydraulic properties determined by inverse modeling of drip infiltrometer experiments extended with pedotransfer functions. Vadose Zone J. 18:180215. doi:10.2136/vzj2018.12.0215

Krause, P., D.P. Boyle, and F. Bäse. 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5:89–97. doi:10.5194/adgeo-5-89-2005

Lamorski, K., Y. Pachepsky, C. Sławiński, and R.T. Walczak. 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Sci. Soc. Am. J. 72:1243–1247. doi:10.2136/sssaj2007.0280N

Liao, K., S. Xu, and Q. Zhu. 2015. Development of ensemble pedotransfer functions for cation exchange capacity of soils of Qingdao in China. Soil Use Manage. 31:483–490. doi:10.1111/sum.12207

Martin, R.T. 1962. Adsorbed water on clay: A review. Clays Clay Miner. 9:28–70. doi:10.1016/B978-1-4831-9842-2.50007-9

McBratney, A.B., B. Minasny, S.R. Cattle, and R.W. Vervoort. 2002. From pedotransfer functions to soil inference systems. Geoderma 109:41–73. doi:10.1016/S0016-7061(02)00139-8

Merdun, H., Ö. Çinar, R. Meral, and M. Apan. 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil Tillage Res. 90:108–116. doi:10.1016/j.still.2005.08.011

Minasny, B., and A. McBratney. 2002. The *neuro-m* method for fitting neural network parametric pedotransfer functions. Soil Sci. Soc. Am. J. 66:352–361. doi:10.2136/sssaj2002.3520

Minasny, B., A.B. McBratney, and K.L. Bristow. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93:225–253. doi:10.1016/S0016-7061(99)00061-0

Møller, A.B., A. Beucher, B.V. Iversen, and M.H. Greve. 2018. Predicting artificially drained areas by means of a selective model ensemble. Geoderma 320:30–42. doi:10.1016/j.geoderma.2018.01.018

Nemes, A., W.J. Rawls, and Y.A. Pachepsky. 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70:327–336. doi:10.2136/sssaj2005.0128

Nemes, A., D.J. Timlin, Y.A. Pachepsky, and W.J. Rawls. 2009. Evaluation of the Rawls (1982) pedotransfer functions for their applicability at the U.S. national scale. Soil Sci. Soc. Am. J. 73:1638–1645. doi:10.2136/sssaj2008.0298

Nguyen, P.M., A. Haghverdi, J. de Pue, Y.-D. Botula, K.V. Le, W. Waegeman, and W.M. Cornelis. 2017. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosyst. Eng. 153:12–27. doi:10.1016/j.biosystemseng.2016.10.013

Nielsen, D.R., and O. Wendroth. 2003. Spatial and temporal statistics: Sampling field soils and their vegetation. Catena Verlag, Reiskirchen, Germany.

Pachepsky, Y., and W.J. Rawls, editors. 2004. Development of pedotransfer functions in soil hydrology. Dev. Soil Sci. 30. Elsevier, Amsterdam.

Patil, N.G., and S.K. Singh. 2016. Pedotransfer functions for estimating soil hydraulic properties: A review. Pedosphere 26:417–430. doi:10.1016/S1002-0160(15)60054-6

Rasmussen, C.E., and C.K.I. Williams. 2006. Gaussian processes for machine learning. MIT Press, Cambridge, MA.

Rawls, W.J., D.L. Brakensiek, and K.E. Saxtonn. 1982. Estimation of soil water properties. Trans. ASAE 25:1316–1320. doi:10.13031/2013.33720

Rawls, W.J., Y.A. Pachepsky, J.C. Ritchie, T.M. Sobecki, and H. Bloodworth. 2003. Effect of soil organic carbon on soil water retention. Geoderma 116:61–76. doi:10.1016/S0016-7061(03)00094-6

Schaap, M.G., and F.J. Leij. 1998a. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil Tillage Res. 47:37–42. doi:10.1016/S0167-1987(98)00070-1

Schaap, M.G., and F.J. Leij. 1998b. Database-related accuracy and uncertainty of pedotransfer functions. Soil Sci. 163:765–779. doi:10.1097/00010694-199810000-00001

Schaap, M.G., F.J. Leij, and M.Th. van Genuchten. 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251:163–176. doi:10.1016/S0022-1694(01)00466-8

Sugiyama, M. 2012. Learning under non-stationarity: Covariate shift adaptation by importance weighting. In: J. Gentle et al., editors, Handbook of computational statistics. Springer, Berlin. p. 927–952. doi:10.1007/978-3-642-21551-3_31

Tomasella, J., M.G. Hodnett, and L. Rossato. 2000. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. Soil Sci. Soc. Am. J. 64:327–338. doi:10.2136/sssaj2000.641327x

Tomasella, J., Y.A. Pachepsky, S. Crestana, and W.J. Rawls. 2003. Comparison of two techniques to develop pedotransfer functions for water retention. Soil Sci. Soc. Am. J. 67:1085–1092. doi:10.2136/sssaj2003.1085

Tóth, B., M. Weynants, A. Nemes, A. Makó, G. Bilas, and G. Tóth. 2015. New generation of hydraulic pedotransfer functions for Europe. Eur. J. Soil Sci. 66:226–238. doi:10.1111/ejss.12192

Tranter, G., A.B. McBratney, and B. Minasny. 2009. Using distance metrics to determine the appropriate domain of pedotransfer function predictions. Geoderma 149:421–425. doi:10.1016/j.geoderma.2009.01.006

Tranter, G., B. Minasny, and A.B. McBratney. 2010. Estimating pedotransfer function prediction limits using fuzzy *k*-means with extragrades. Soil Sci. Soc. Am. J. 74:1967–1975. doi:10.2136/sssaj2009.0106

van den Berg, M., E. Klamt, L.P. Van Reeuwijk, and W.G. Sombroek. 1997. Pedotransfer functions for the estimation of moisture retention characteristics of Ferralsols and related soils. Geoderma 78:161–180. doi:10.1016/S0016-7061(97)00045-1

van Genuchten, M.Th. 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44:892–898. doi:10.2136/sssaj1980.03615995004400050002x

van Genuchten, M.Th., F.J. Leij, and S.R. Yates. 1991. The RETC code for quantifying the hydraulic functions of unsaturated soils. USEPA, Robert S. Kerr Environ. Res. Lab., Ada, OK.

van Looy, K., J. Bouma, M. Herbst, J. Koestel, B. Minasny, U. Mishra, et al. 2017. Pedotransfer functions in Earth system science: Challenges and perspectives. Reb. Geophys. 55:1199–1256. doi:10.1002/2017RG000581

Vereecken, H., J. Maes, J. Feyen, and P. Darius. 1989. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. Soil Sci. 148:389–403. doi:10.1097/00010694-198912000-00001

Wösten, J.H.M., Y.A. Pachepsky, and W.J. Rawls. 2001. Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251:123–150. doi:10.1016/S0022-1694(01)00464-4

Ye, M., R. Khaleel, M.G. Schaap, and J. Zhu. 2007. Simulation of field injection experiments in heterogeneous unsaturated media using cokriging and artificial neural network. Water Resour. Res. 43:W07413. doi:10.1029/2006WR005030

Zhang, C., and Y. Ma, editors. 2012. Ensemble machine learning: Methods and applications. Springer, New York. doi:10.1007/978-1-4419-9326-7

Zhao, C., M. Shao, X. Jia, M. Nasir, and C. Zhang. 2016. Using pedotransfer functions to estimate soil hydraulic conductivity in the Loess Plateau of China. Catena 143:1–6. doi:10.1016/j.catena.2016.03.037

Zhao, C., M. Shao, X. Jia, and Y. Zhu. 2017. Estimation of spatial variability of soil water storage along the south–north transect on China's Loess Plateau using the state-space approach. J. Soils Sediments 17:1009–1020. doi:10.1007/s11368-016-1626-8