Hindawi Publishing Corporation
Applied and Environmental Soil Science
Volume 2015, Article ID 535216, 16 pages
http://dx.doi.org/10.1155/2015/535216

Research Article

Artificial Neural Networks for Estimating Soil Water Retention Curve Using Fitted and Measured Data

Tirzah Moreira de Melo and Olavo Correa Pedrollo

Institute of Hydraulic Researches, Federal University of Rio Grande do Sul (UFRGS), Bento Gonçalves Avenue 9500, P.O. Box 15029, 91501-970 Porto Alegre, RS, Brazil

Correspondence should be addressed to Tirzah Moreira de Melo; tirzahmelo@hotmail.com

Received 7 August 2014; Accepted 24 February 2015

Academic Editor: Keith Smettem

Artificial neural networks for estimating the soil water retention curve have been developed considering measured data and require a large quantity of soil samples because only retention curve data obtained for the same set of matric potentials can be used. In order to preclude this drawback, we present two ANN models which tested the performance of ANNs trained with fitted water contents data. These models were compared to a recent new ANN approach for predicting water retention curve, the pseudocontinuous pedotransfer functions (PTFs), which is also an attempt to deal with limited data. Additionally, a sensitivity analysis was carried out to verify the influence of each input parameter on each output. Results showed that fitted ANNs provided similar statistical indexes in predicting water contents to those obtained by the pseudocontinuous method. Sensitivity analysis revealed that bulk density and porosity are the most important parameters for predicting water contents in wet regime, whereas sand and clay contents are more significant in drier conditions. The sensitivity analysis for the pseudocontinuous method demonstrated that the natural logarithm of the matric potential became the most important parameter, and the influences of all other inputs were reduced to be not relevant, except the bulk density.

1. Introduction

Modelling water flow and solute transport in vadose zone is generally done by means of Richard's equation and convection-dispersion equation (CDE), respectively, which in turn require some crucial soil information, such as unsaturated soil hydraulic properties or functions and the soil water retention curve. In agriculture applications, knowledge of these properties is needed, for instance, when dealing with irrigation and drainage management, analysis of biological reactions, plant activity, and stream water chemistry [1]. Thus, the prediction accuracy of these models hinges upon the quality of the model parameters.

Nevertheless, direct measurements of these properties at any grid scale are labour intensive, expensive, and time-consuming. The more reliable the analysis of the variability of these spatially distributed parameters, the larger the number of samples to be collected [2]. In addition, the relations between soil properties and soil-water processes are believed to be highly nonlinear, and in most of the cases they cannot be easily modelled by simple mathematical formulations or even by complex models, which need a high number of input parameters.

Since all these factors make direct measurements of hydraulic properties or the use of some models impractical, pedotransfer functions (PTFs) have been widely developed for the last decades in order to estimate the hydraulic conductivity function and the soil water retention curve from other more easily measurable soil properties, such as texture, bulk density, organic matter, porosity, and particle-size distribution [3–13]. In some cases, when changing from a field scale-up to the scale of the catchment, the use of pedotransfer functions could be the only way to apply hydrological models [7].

In the context of soil water processes, PTFs can be categorized into three main groups: class, point, and parametric PTFs. The first type calculates hydraulic properties for a texture class by assuming that similar soils have similar hydraulic properties [14]. Point PTFs predict soil water contents at specific matric potentials whereas parametric PTFs estimate the parameters of a soil hydraulic model *a priori* defined, such as the Brooks and Corey [15] or the van Genuchten [16] equations. Among several well-known PTFs, irrespective of their abovementioned classification, artificial neural networks (ANNs) have become the most frequently used [4].

ANN is an attempt to develop a model which works similarly to the human brain and is analogous to the biological function of learning and memorizing, comprising a densely network of connections between input data, neurons disposed in different hidden layers with parameters to be fitted, and output data. Its main advantage is that it does not need a previous knowledge of the relations between input and output data; that is, in our case, no consideration of the internal geologic or hydraulic parameters is required [1]. On the other hand, the main drawback is that ANNs are data hungry, so that their development is highly associated with the existence of large soil hydraulic properties databases [17]. An additional disadvantage is the "black-box" nature of ANN, which makes it difficult to go beyond a strictly empirical model.

More recently, Haghverdi et al. [4] have developed a new ANN approach called pseudocontinuous model, which is capable of predicting water content at any desirable matric potential without the need of any specific equation, by simply adding the matric potential as input parameter (sand, silt, clay, bulk density, and organic matter being the other input parameters) and water content being the only output neuron. The major contribution of this new ANN is that it can be used when a limited number of soil samples are available, resulting in the fact that datasets more heterogeneous could be employed.

Additionally, it is extremely useful if soil water retention curves measured for distinct matric potentials are added in the process of training a neural network and not only those curves with the same matric potentials. This fact increases the number of examples and allows for the development of new PTFs based on limited data.

Another way to solve the problem of having only few soil samples is previously fitting the retention curve for a wider range of matric potentials and then training the ANN with more information about the soils. In comparison with the work of Haghverdi et al. [4], this approach has the advantage of permitting domain extrapolations, since the ANN model will be able to estimate water contents for matric potentials other than those which were sampled.

The proposed methodology could be easily validated by modelling and analysing the quality of the results. An option of relative assessment of the ANN models' quality is the comparison of some statistics with those obtained with other approaches, such as the model of Haghverdi et al. [4].

Therefore, considering the advantages of this new approach and the aforesaid considerations, the present paper focuses on (1) deriving one ANN, a pseudocontinuous reference model, for predicting soil water retention curve based on measured water contents; (2) deriving two ANNs based on fitted water contents to be validated in comparison with the pseudocontinuous one; (3) carrying out a sensitivity analysis of the importance of the input parameters upon the outputs.

Table 1: Descriptive statistics of soil properties assumed as input and output parameters in the developed ANNs (all 228 soil samples).

Soil property	Input parameters			
	Min	Max	Mean	SD
Bd (g cm^{-3})	0.46	1.95	1.41	0.26
Pd (g cm^{-3})	1.98	2.93	2.63	0.09
P (cm^3 cm^{-3})	0.26	0.92	0.47	0.10
Sand (%)	0.00	98.90	26.30	25.58
Silt (%)	1.10	95.60	40.84	18.79
Clay (%)	0.00	87.60	32.82	21.49
ln(h) (cm)	0.00	15.62	5.20	2.50
Matric potentials (cm)	**Output parameters**			
	Min	Max	Mean	SD
Measured	0.008	0.838	0.314	0.140
0	0.139	0.837	0.449	0.117
−10	0.139	0.826	0.427	0.108
−20	0.139	0.789	0.408	0.106
−30	0.108	0.749	0.391	0.108
−50	0.066	0.707	0.364	0.113
−100	0.043	0.631	0.329	0.118
−200	0.037	0.597	0.299	0.118
−500	0.010	0.576	0.265	0.112
−1000	0.003	0.568	0.241	0.108

The reference model comprises the pseudocontinuous approach of Haghverdi et al. [4] but considers input parameters different from those employed by these authors. The first model tested consists of a multilayered point ANN with several output neurons, each one corresponding to a distinct matric potential, and the second considers different ANNs of only one output neuron, being one ANN for each specific matric potential. Thus, firstly it is of interest to validate the two models when compared to the pseudocontinuous method and then choose which ANN structures provided better performances.

2. Materials and Methods

2.1. Soil Samples. A total of 228 soil samples were selected from the international UNSODA database (version 2.0), which contains soil data collected from different parts of the world, covering a wide range of soil types and characteristics. A detailed description of this database can be found in Leij et al. [18] and Nemes et al. [19].

Corresponding data on bulk density (Bd), total porosity (P), particle density (Pd), sand, silt, and clay percentages, and the soil water retention curve measured points were selected and, thus, the first six properties were assigned as input parameters and the soil moisture of the retention curve was assigned as the output parameter for the three developed ANNs. Table 1 briefly presents the descriptive statistics of selected input and output parameters (all 228 soil samples). For some soil samples, missing total porosity was calculated according to $P = 1 - \mathrm{Bd}/\mathrm{Pd}$. The full database was divided into

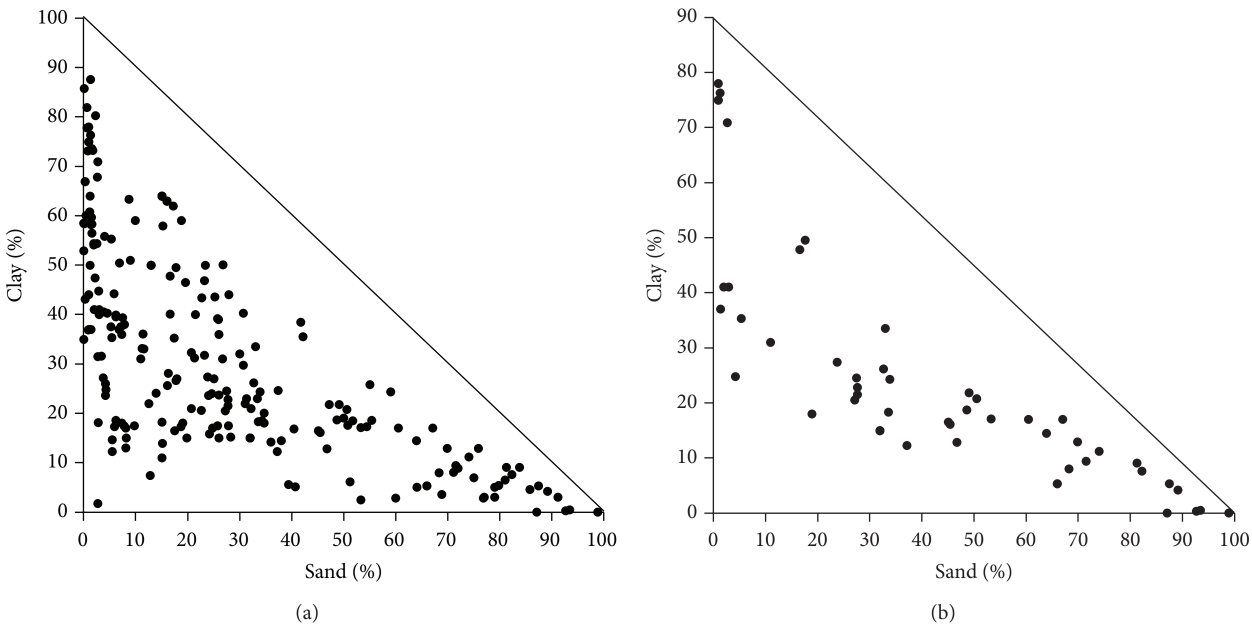

Figure 1: Soil texture of all selected soils (a) and verification dataset (b). Clay (<0.002 mm) and silt (0.002–0.05 mm).

three subsets to run the ANNs, being 137 for training, 40 for validation, and 51 for verification data (see Section 2.3).

We firstly tested the influence of how data is selected for each subset. This process was taken into consideration by randomly dividing the full dataset in three subsets and then analyzing the error estimation of the water content (outputs). We forced that the verification dataset had at least 60% of the full dataset, in order to guarantee greater chance to include in this dataset the maximum and minimum values of each input parameter. The final configuration was selected after comparing the errors of each test. Figures 1 and 2 show the textural distribution for all 228 soils and for the verification dataset (Figure 1(b)), from which it is possible to notice that a wide range of soil textures is represented by the whole dataset.

The performances of the proposed ANNs were tested confronting the use of measured water contents as output parameter by the pseudocontinuous ANN against the assumption of fitted water content data given by the van Genuchten [16] model in the other two ANNs, given as

$$\theta(h) = \theta_r + \frac{\theta_s - \theta_r}{\left(1 + |\alpha h|^n\right)^m}, \qquad (1)$$

where $\theta(h)$ is the water content (cm^3 cm^{-3}) at matric potential h (cm); θ_s and θ_r are the saturated and residual water contents (cm^3 cm^{-3}), respectively; $n(-)$ is the curve shape factor which controls the steepness of the S-shaped retention curve; m is also an empirical shape factor (−) related to n by $m = 1-(1/n)$; and α is assumed to be related to the inverse of air entry suction (cm^{-1}), indicating typical conditions of sands for large values while very small negative matric potentials empty pores creating a relatively large change in water content [14].

Therefore, for noncontinuous models, nine matric potentials were fixed at 0, −10, −20, −30, −50, −100, −200, −500, and −1000 cm. Water contents at these suctions were estimated by fitting the van Genuchten model parameters from measured retention curve data, so that the output parameters could be defined for the same matric potentials. The limit of −1000 cm was chosen to avoid great extrapolations for some soil samples with higher observed matric potentials. Fitting was carried out by the RETC software [20].

Fitted values of water contents were also employed and tested by Koekkoek and Booltink [21] for Dutch soils and Santra and Das [22] fitted the van Genuchten parameters in order to take the porosity equal to the saturated water content as an input parameter in their PTFs.

2.2. Artificial Neural Networks (ANNs). A model of artificial neural network usually consists of several layers, input, hidden, and output ones, connected via parameters and called, in analogy with the natural neural networks, the synaptic weights. The training provides the ability of approximating the relationship between the inputs and the corresponding outputs. After having been duly submitted to a training process, the ANN is able to generalize the learned relation to other samples in the same domain.

The number of hidden layers in a multilayer neural network can vary, depending on the complexity and nature of the problem to be modelled. The main function of this layer is to allow the network to capture both weak and strong nonlinear relationships between inputs and outputs. In the present context, we are interested in translating basic information from the input layer, for example, particle-size distribution, bulk density, and so forth, into, not easily measurable data of the output layer, the soil water retention curve (SWRC).

It is necessary to go through a "learning" or training process involving the adjustment of the synaptic connections (weights and biases). The most widely used training scheme is

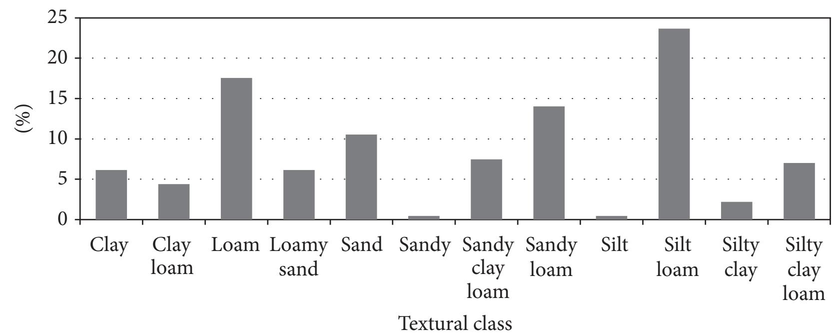

Figure 2: Textural composition of the dataset.

the back-propagation algorithm, which basically involves two steps: the first one is the forward phase, when the activations of the neurons are propagated from inputs to outputs, and the second one consists of the back-propagation [23] to the inner layers of the errors between the observed and estimated values in the output layer and modifying the weights and bias coefficients through the use of the delta rule [24], given, respectively, by the following equations:

$$e_s = T_s - O_s, \tag{2}$$

where e_s is the error of the output layer and T_s and O_s are, respectively, the target and the calculated outputs for this layer;

$$e_h = \sum \left(W_s e_s \delta_s\right), \tag{3}$$

where e_h is the error of the hidden layer and W_s, e_s, and δ_s are, respectively, the weights, the errors, and the derivatives of the activation function of the next layer;

$$W_{k+1} = W_k + \tau e_k \delta_k P_k, \tag{4}$$

where, to a generic layer and considering successive cycles (each iteration over all the patterns) of the training sample, W_k are the resulting weights; τ is the learning rate; e_k is the vector of errors in the output of this layer; δ_k is the derivative of the activation function; and P_k are the inputs of this layer in the cycle k.

The forward phase starts based on a random "initial condition" of the weights and biases connecting all layers, being the net input in each layer through the ANN given by

$$\text{net}_j = \sum_{i=1}^{n} w_{ij} P_i + b_j, \tag{5}$$

where net_j is the net input (sum of the weighted input received from the preceding layer with n neurons) of the neuron j, P_i represents the output of the ith neuron in the previous layer, w_{ij} is the weight between the ith neuron in the previous layer and the j neuron, and b_j is the bias coefficient of the neuron j.

The response O_j given by hidden or output neurons is calculated by an activation function, providing the input to the next layer. In our case, the log-sigmoid transfer function was considered for the hidden and output layers:

$$O_j = \frac{1}{1 + e^{-\text{net}_j}}. \tag{6}$$

For all the three ANNs developed in this study, a multilayer perceptron neural network was used, which trains the networks employing a back-propagation algorithm. The initial coefficients w and b were randomly initialized following a normal distribution.

2.3. Models. The first model applied is called the reference model (ANN_o) and is analogous to the pseudocontinuous approach introduced by Haghverdi et al. [4], except for the use of different input parameters as shown in Table 1. The method consists of adding the natural logarithm of matric potential as input parameter, as depicted in Figure 3. As a consequence, only one output neuron is needed, corresponding to the measured water contents at the whole range of the considered matric potentials. The approach could be seen as a sequential adjustment of the soil water retention curves, whereas point ANNs try to fit these curves in parallel using several output neurons. The impact on the input parameters matrix is that for each sequence of matric potentials (each SWRC) all other input parameters are repeated, resulting in a higher number of available data for adjusting the same number of outputs. That is the reason why this method is believed to be useful for developing ANN-based PTFs under a limited number of soil samples, chiefly because all measured SWRC points could be used in spite of their corresponding distinct matric potentials.

The other two models were developed to test their performances of using fitted data against the use of measured data (ANN_o). The first model (ANN 1) consists of a point neural network of nine output neurons, each one predicting the water contents related to a different matric potential as presented in Table 1, while the first six input parameters in this table were used.

The ANN 2 model, actually, is a set of different ANNs of only one output neuron. Hence, for each ANN_i, the same six input parameters as in ANN 1 model were used, but water contents of only one matric potential was considered at a time. The structures of all ANNs are presented in Figure 3.

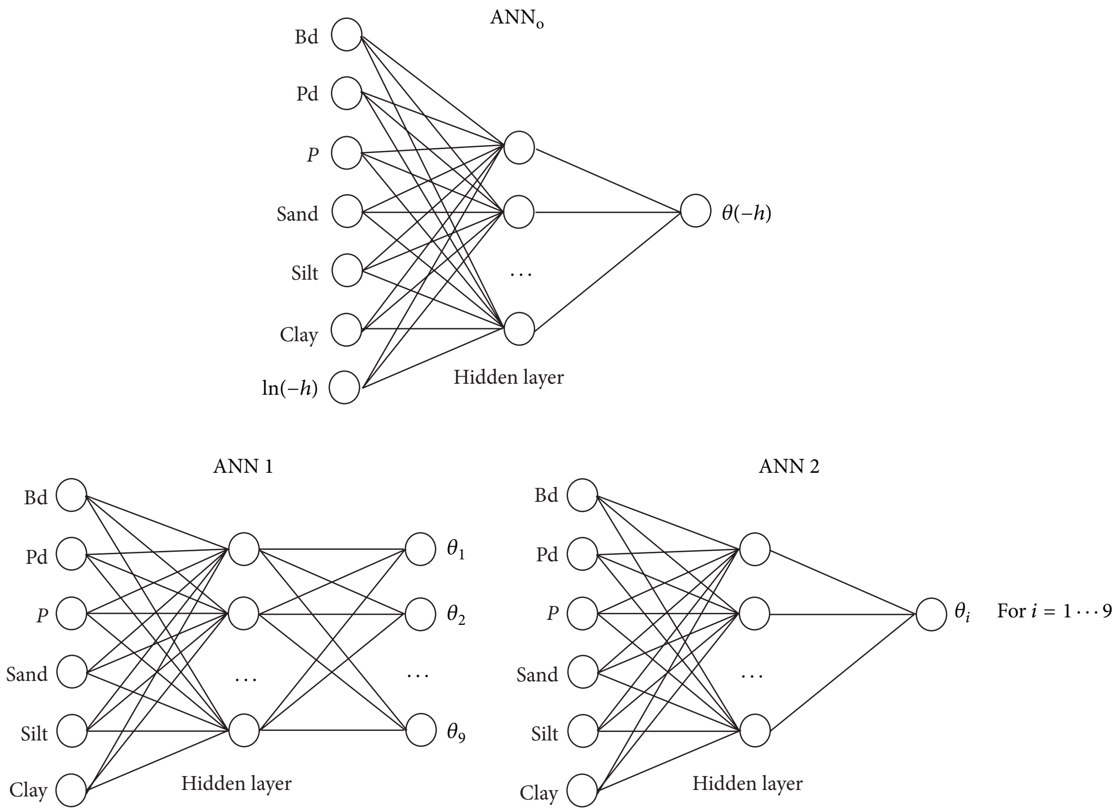

FIGURE 3: Models diagrams of the ANNs developed in this study. Bulk density (Bd), particle density (Pd), total porosity (P), and sand, silt, and clay percentages are the common input predictors of the three models. In ANNs 1 and 2, θ_i (for $i = 1\cdots 9$) are the water content at the matric potentials 0, −10, −20, −30, −50, −100, −200, −500, and −1000 cm, respectively. In ANN 3, the matric potential ($-h$) was considered as an additional predictor for estimating the corresponding water content $\theta(-h)$.

Direct utilization of the input parameters does not provide good estimates of water contents because there is a significant difference in the magnitude of inputs. Thus, a preliminary linear transformation was carried out, resulting in all input properties varying between 0 and 1. The same process was performed with the outputs and then the values were back-transformed to calculate the error estimates.

To prevent data from overfitting, the technique of cross-validation or "early stopping" is used. Such a method consists of monitoring the error on the validation dataset; that is, when the validation error increases for a specified iteration, the training process is terminated and the weights and bias coefficients are assumed to generate the outputs.

The algorithms of the developed ANNs (including the reference model) executed 20 different trainings with cross-validation and simultaneously tested the number of neurons in the hidden layer. Each training performed the number of 1000 iterations, assuming as "the best model" the model corresponding to the best combination of number of neurons, training realisation, and minimum validation error. The number of 1000 was chosen after observing that training usually stopped before such quantity of iterations, while several trainings were carried out in order to avoid the influences of random initial weights and biases, which could prematurely stop in a local minimum on the error surface.

The error surface corresponds to the objective function to be minimized during the network training, given by

$$E = \sum_{s=1}^{N} \sum_{n=1}^{m} \left(T_{sn} - O_{sn}\right)^2, \tag{7}$$

where N is the number of input and output samples; m is the number of neurons in the output layer; T_{sn} is the target value of the nth neuron for the sth sample; and O_{sn} is the output of the nth neuron for the sth sample. All ANN constructions and calculations were conducted in the MatLab programming language written by the authors.

2.4. Statistical Performance Criteria of ANNs. ANN performances were assessed in terms of the agreement between the predicted and measured water contents for each matric potential. Different statistical indexes were used so that some distinct characteristics of the models could be evaluated and interpreted. The chosen indexes were the root mean square of error (RMSE), the coefficient of determination (R^2), and the geometric mean error ratio (GMER), expressed as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{j}\left(O_i - P_i\right)^2}{j}}$$
$$R^2 = 1 - \frac{\sum_{i=1}^{j}\left(O_i - P_i\right)^2}{\sum_{i=1}^{j}\left(O_i - \overline{O}\right)^2} \tag{8}$$
$$\mathrm{GMER} = \exp\left[\frac{1}{j}\sum_{i=1}^{j}\ln\left(\frac{P_i}{O_i}\right)\right],$$

where j is the number of samples; O_i and P_i are the observed and predicted water contents, respectively; and $\overline{O}$ is the mean of the observed data.

The RMSE is a measure for the accuracy of the estimations in terms of standard deviations, whilst R^2 indicates how much of the variance between measured and estimated water contents could be explained by the ANNs. The last index (GMER) was selected instead of using the correlation coefficient (R) because R index only indicates that the scatter plot falls almost along a straight line with positive or negative slope, but nothing is explained about the inclination of this line; that is, measured data could be any multiple of estimated data and R keeps being approximately 1. Diversely, GMER equal to 1 corresponds to an exact matching between measured and predicted data [25], and values less or greater than 1 indicate under- or overestimation of the measured water contents. The means of predicted and measured data were also compared.

Thus, the statistical indexes of $\mathrm{ANN_o}$ will serve as reference to assess the performance of the other two ANN models. After proving their capability of predicting water contents, ANN 1 and ANN 2 will be confronted in terms of their structures, deciding which one is more suitable.

2.5. Sensitivity Analysis. In order to assess the importance of each input parameter to the overall model performances of individual output, the sensitivity analyses were conducted. Many researchers [4–6, 9, 10, 22] have usually proposed the combination of single-parameter influences for assessing their importance.

The applied method is an improvement of the classical stepwise method and consists of replacing step by step one input predictor by its mean value and assessing the change in any statistical index. Thus, this is an approach to evaluate the general influence of each input parameter on the model performance, detecting which are the most important to be considered or rejected. The accuracy and precision of this method, among others, were evaluated by Gevrey et al. [26] and Olden et al. [27] and correspond to one of the most suitable approaches suggested by these authors for more accurate quantifying of the importance of variables.

3. Results and Discussions

3.1. Model Performances. The statistical indexes for assessing ANN performances are presented in Table 2 and refer to the process of verification only. The results obtained for the reference model ($\mathrm{ANN_o}$) provided RMSE of 0.088 $\mathrm{cm^3\ cm^{-3}}$, which could be considered satisfactory when compared to other studies. The R^2 index revealed that most of the variance of the phenomenon could be captured by the model. On the other hand, the GMER index showed a significant and meaningful water contents underestimation expressed by a value lower than 1. This fact can be confirmed by the scatter plots presented in Figure 4. These graphics refer to predicted and fitted/measured water contents of all matric potentials of the verification dataset.

For ANN 1 and ANN 2, the statistics are shown for each matric potential and the means were taken so that they could be compared to $\mathrm{ANN_o}$. It is important to mention that all results presented herein for fitted data refers to fittings with regression coefficients R greater than 0.99. RMSE values varied from 0.050 to 0.083 and from 0.054 to 0.098 $\mathrm{cm^3\ cm^{-3}}$ for ANN 1 and ANN 2, respectively, being the minimum values corresponding to the matric potentials −50 cm (ANN 1) and −10 cm (ANN 2) and the maximum values to the matric potentials −500 cm (ANN 1) and −200 cm (ANN 2). These RMSE values are comparable to those found by Puhlmann and von Wilpert [3], Liao et al. [28], and Baker and Ellison [17] and are much better than those obtained by Koekkoek and Booltink [21]. These last authors also applied fitted water contents to their ANN models, but values of RMSE higher than 2.64 $\mathrm{cm^3\ cm^{-3}}$ were found.

When the results of ANN 1 and ANN 2 are compared to those obtained by the method of Haghverdi et al. [4], an opposite behaviour was found, since they achieved lower RMSE values at lower (more negative) suctions.

One possibility for this fact is that porosity was also added as predictor in the present study and could be an important parameter for accounting for the greater variation of water contents in wet regime, since this section is hardly governed by soil structure.

Mermoud and Xu [29] confirmed it mentioning that soil structure is crucial in characterizing hydraulic behaviour in macropore flow region, whilst the flow in micropores is highly influenced by texture.

For all matric potentials, values of GMER near to 1 were obtained, which means that good matching was provided by both ANN 1 and ANN 2. Again, better performances resulted from the wet section of the retention curve. This index also reflects that ANN 1 and ANN 2 slightly overestimate target values (GMER > 1), while $\mathrm{ANN_o}$ underestimates (GMER < 1) them. This fact could be also confirmed in the scatter plots presented in Figure 4.

Finally, analysing all ANNs together and assuming the mean values of the statistic coefficients of ANN 1 and ANN 2 to compare each of them with $\mathrm{ANN_o}$ (Table 2), similar performances were found, except for the higher RMSE index obtained by $\mathrm{ANN_o}$. Comparing ANN 1 and ANN 2, the best performance seems to be the one given by the neural network whose structure trains all matric potentials together (ANN

Table 2: Statistical analysis of artificial neural networks developed in this study for predicting the soil water retention curve.

Index	Matric potentials (−cm) 0 ANN 1	0 ANN 2	10 ANN 1	10 ANN 2	20 ANN 1	20 ANN 2	30 ANN 1	30 ANN 2	50 ANN 1	50 ANN 2	100 ANN 1	100 ANN 2	200 ANN 1	200 ANN 2	500 ANN 1	500 ANN 2	1000 ANN 1	1000 ANN 2	Mean ANN 1	Mean ANN 2	ANN_o
RMSE	0.064	0.062	0.057	0.054	0.056	0.055	0.051	0.055	0.050	0.060	0.069	0.075	0.082	0.098	0.083	0.087	0.078	0.084	0.066	0.070	0.088
R^2	0.697	0.712	0.727	0.756	0.755	0.763	0.800	0.773	0.818	0.740	0.699	0.645	0.606	0.433	0.559	0.508	0.543	0.479	0.689	0.645	0.640
GMER	1.062	1.066	1.041	1.036	1.032	1.026	1.027	1.035	1.025	1.055	1.077	1.106	1.136	1.076	1.228	1.278	1.272	1.346	1.100	1.114	0.923
Target mean	0.428		0.407		0.388		0.368		0.339		0.299		0.268		0.236		0.216		0.328[a]	0.328[a]	0.331[b]
Estimated mean	0.448	0.444	0.414	0.412	0.388	0.387	0.365	0.367	0.334	0.339	0.300	0.303	0.271	0.247	0.238	0.242	0.213	0.219	0.330	0.329	0.273

[a] In this case, target mean refers to water contents fitted by van Genuchten parameters for each matric potential. [b] The mean value of observed water contents.

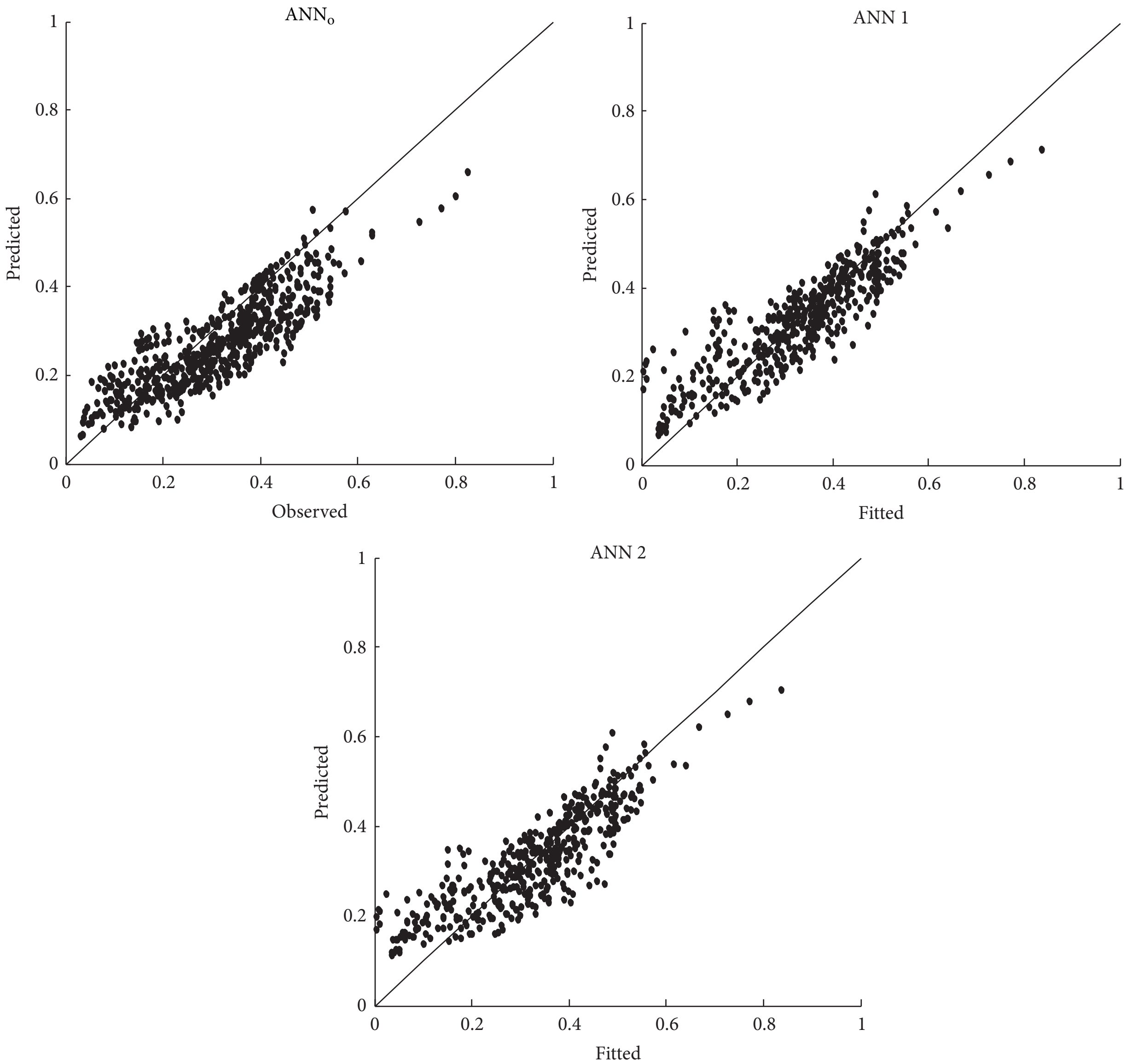

Figure 4: Scatter plots of fitted (ANN 1 and ANN 2 models) and observed (ANN 3) *versus* predicted water contents (cm^3 cm^{-3}) for all matric potentials.

1) and, probably, may be more capable of capturing more accurately the behaviour of the retention curve even for drier conditions. Yet, the pseudocontinuous method developed by Haghverdi et al. [4] was only slightly better than the other point and parametric PTFs they have tested, indicating that the results in this study are in concordance with those presented by these authors.

Performances of ANN 1 and ANN 2 models are very similar and, in general, ANN 2 model performed slightly better than ANN 1 for matric potentials corresponding to wet conditions as reflected by all statistics. Additionally, it could be noted also that both models can more accurately explain the variability of the wet regime of the retention curve than for dry water contents, according to the index R^2.

The water content overestimation in drier condition happened for soils samples with higher sand percentages, whose water retention curves were almost vertical, abruptly decaying. Yet, for these soils, the measured points of the retention curve did not contemplate the whole range of matric potentials considered for fitting (no measure points were available for conditions drier than −200 cm), so larger errors resulted for these suctions. For sand soils with other measured points beside −200 cm, the overestimation was lower than for the former (e.g., for the sand soil code 2562, there were measured points till the range of about −700 cm). Additionally, as it can be seen in Figure 2, sandy soils have only few samples, making it more difficult to model this type of soil. Thus, one reason for overestimation is the

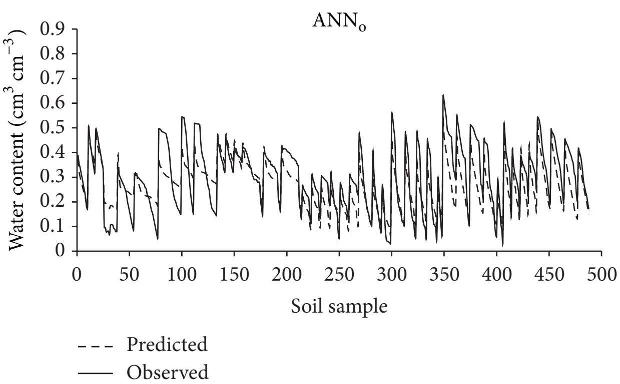

Figure 5: Observed and predicted water contents obtained by ANN$_o$ for soils from the verification dataset.

underrepresentation of some classes of soil textures in the verification dataset (Figure 1). One possibility for precluding this drawback is selecting other subsets for training, validation, and verification processes, since Figure 4 refers to soil samples of verification dataset only.

Figures 5, 6, and 7 present predicted water contents given by the developed ANNs, from where we can verify a good visual matching of the predicted and target (fitted or measured) water contents. In Figure 5, each segment of ups and downs corresponds to a different retention curve.

Worse performances for soil samples from 25 to 200 in ANN$_o$ probably are due to the fact that most of the soils from this range contain larger percentages of sand (sand and sandy loam soils) and even because measured water contents contemplate a smaller range of matric potentials. For these soils, an abrupt decay in water contents is observed and none of the ANNs was able to capture it properly. Notice that the range of soil samples from 25 to 200 in ANN$_o$ is equivalent to the range from 7 to 16 in ANN 1 and ANN 2.

In order to illustrate the ANN performances in estimating the soil water retention curve, Figure 8 presents some curves obtained for all soil types included in the verification dataset. It is clear that all ANNs performed well near saturation. From most of the cases, ANN 1 gave the best agreement with observed data, whereas ANN 2 displays a surprising behaviour for matric potentials between −200 and −400 cm, for which the graphic suffers a small curvature.

One reason for that behaviour could be the lack of information for characterizing the input parameters that more markedly affect water contents associated with these suctions, once the network has trained each matric potential singly. Therefore, this kind of ANN should not be a good choice. This fact could be better understood by performing the sensitivity analysis of each input predictors on each output, as discussed in the next section.

The ANN$_o$ model reproduced the curve shapes very similar to ANN 1, sometimes overestimating it and sometimes underestimating it. However, it is extremely difficult to establish a pattern from inputs followed by good fittings of outputs. Some general remarks that could be drawn from the results are as follows: (1) both high and low silt percentages provided good estimates; hence, it is believed that silt could be of lower importance than the other parameters; (2) the best agreement was found for very low sand percentages and high percentages of silt and clay; (3) the worst results were given by the combination of high percentages of sand and very low percentages of clay. It is also worth noting that these conclusions are in concordance with those stated by Romano and Santini [30], who argued that the largest deviations between fitted and PTF-estimated water contents are mainly in consequence of those samples showing lower sand content (about 11%) or low values of bulk density (about 1.19 g/cm^3). In our study, worse deviations were found for bulk densities lower than 1,34 g/cm^3. The authors also mentioned that it is not the influence of these properties alone that are important in causing poor results but their simultaneous occurrence. Additionally, some poor results could be associated with the effect of underrepresentation of some classes of soil textures in the dataset, such as high sand and low clay percentages.

Comparisons of target mean values and estimated mean values (report to Table 2) are in good agreement for all ANNs, showing that all developed models are able to reproduce the central tendency of target data, being the greatest deviation from ANN$_o$ model. All these results reflect that observed data is not necessarily the only way to train an ANN, once we not only are interested in predicting the individual water contents but also are trying to develop PTF-based models which are efficient in accurately predicting the shape of the retention curve. In the present study, this characteristic was more effectively achieved by ANNs trained with fitted data. These ANNs could be seen as the inverse process of a parametric PTF, when we firstly fit any model parameters and train the PTF to latter estimate water contents using such predefined equation. In both cases, water contents are being predicted based on fitted parameters for deriving the soil water retention curve.

3.2. Sensitivity Analysis. Evaluation of PTFs is generally based on statistical criteria that compare predicted results and experimental data, but nothing is mentioned about what is contributing to the performance of the models or about their predictive behaviour in estimating hydraulic properties. Yet, when any method of sensitivity analysis is carried out, the most common technique consists in simply constructing several PTFs with different combinations of input parameters and checking the change in a specific statistical index, which is labour- and time-consuming. In order to preclude these drawbacks, we employed the improved stepwise method aforementioned.

In Table 3, the results expressed in terms of the percentage variation of RMSE index are presented. In this table, "model's RMSE" is the index of prediction of each output (matric potentials) given by ANNs, while the "RMSE" for each parameter reports to the resulting performance when such an input is replaced by its mean. Thus, positive Var (%) indicates that this ANN perturbation was reflected by an increase of the error, consequently, by a decrease in predictive performance, and vice versa.

For ANN 1 and ANN 2 models, wet regime is more strongly influenced by bulk density and porosity, while

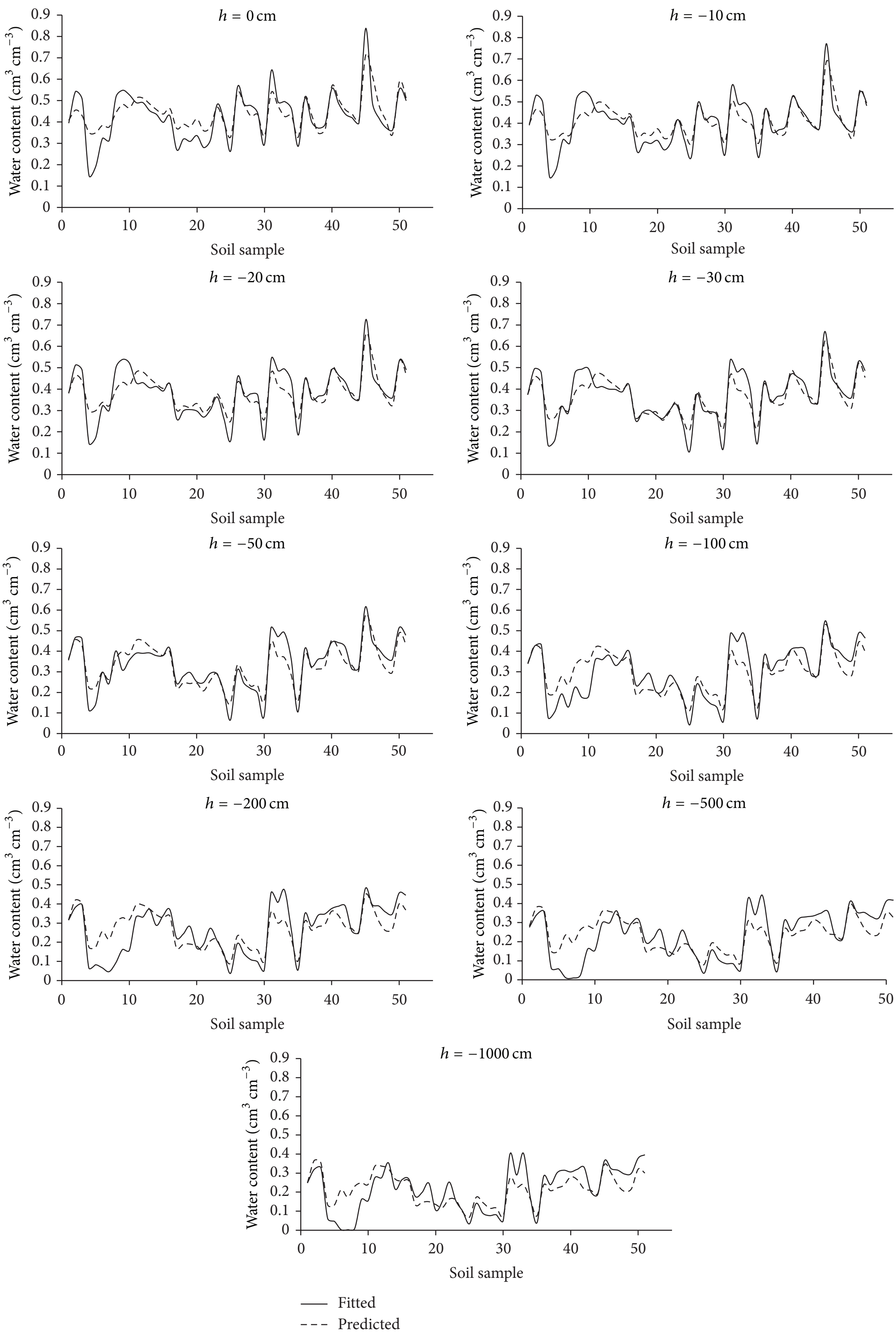

Figure 6: Fitted and predicted water contents obtained by ANN 1 model and van Genuchten parameters, respectively, where h is the matric potential.

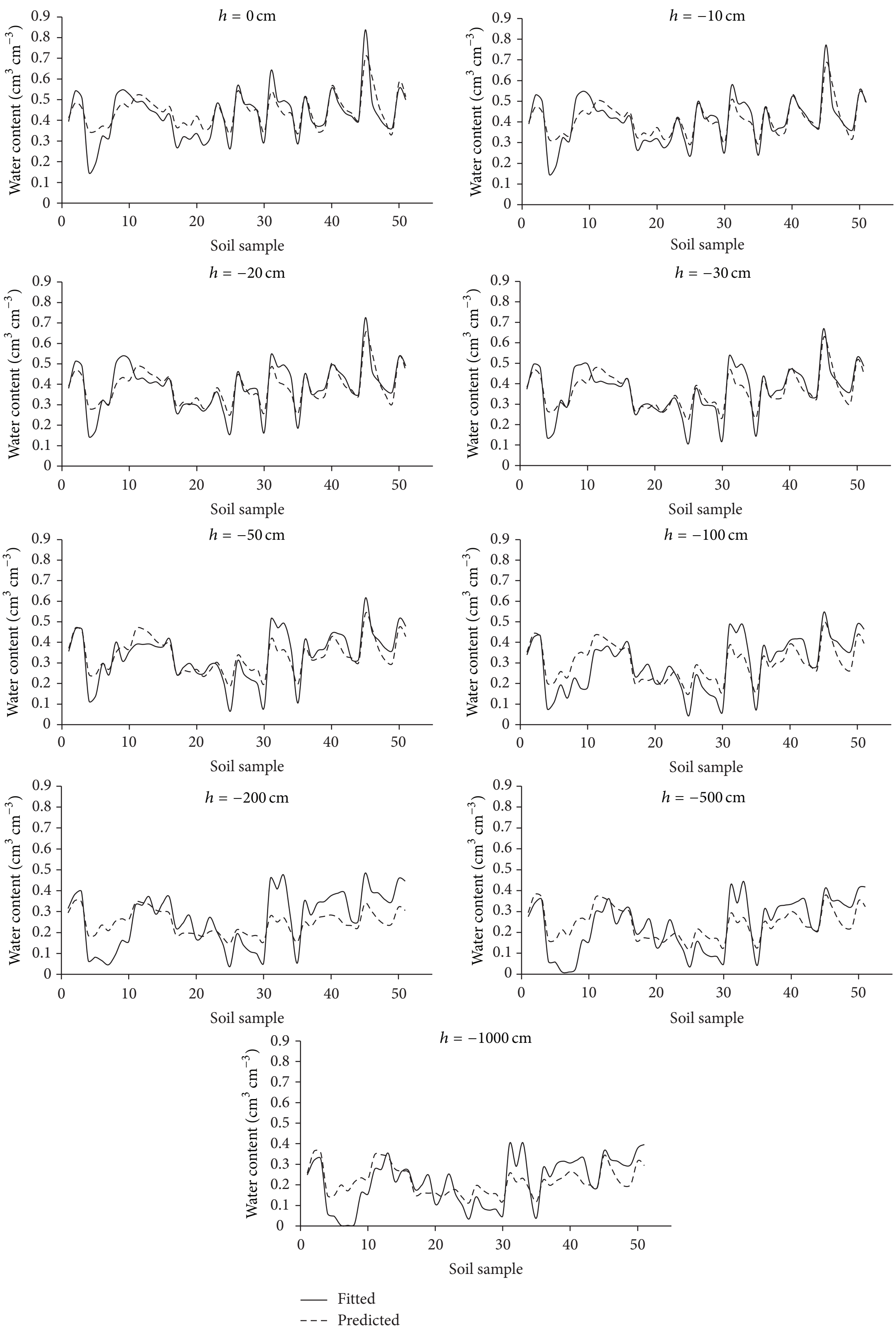

Figure 7: Fitted and predicted water contents obtained by ANN 2 model and van Genuchten parameters, respectively, where h is the matric potential.

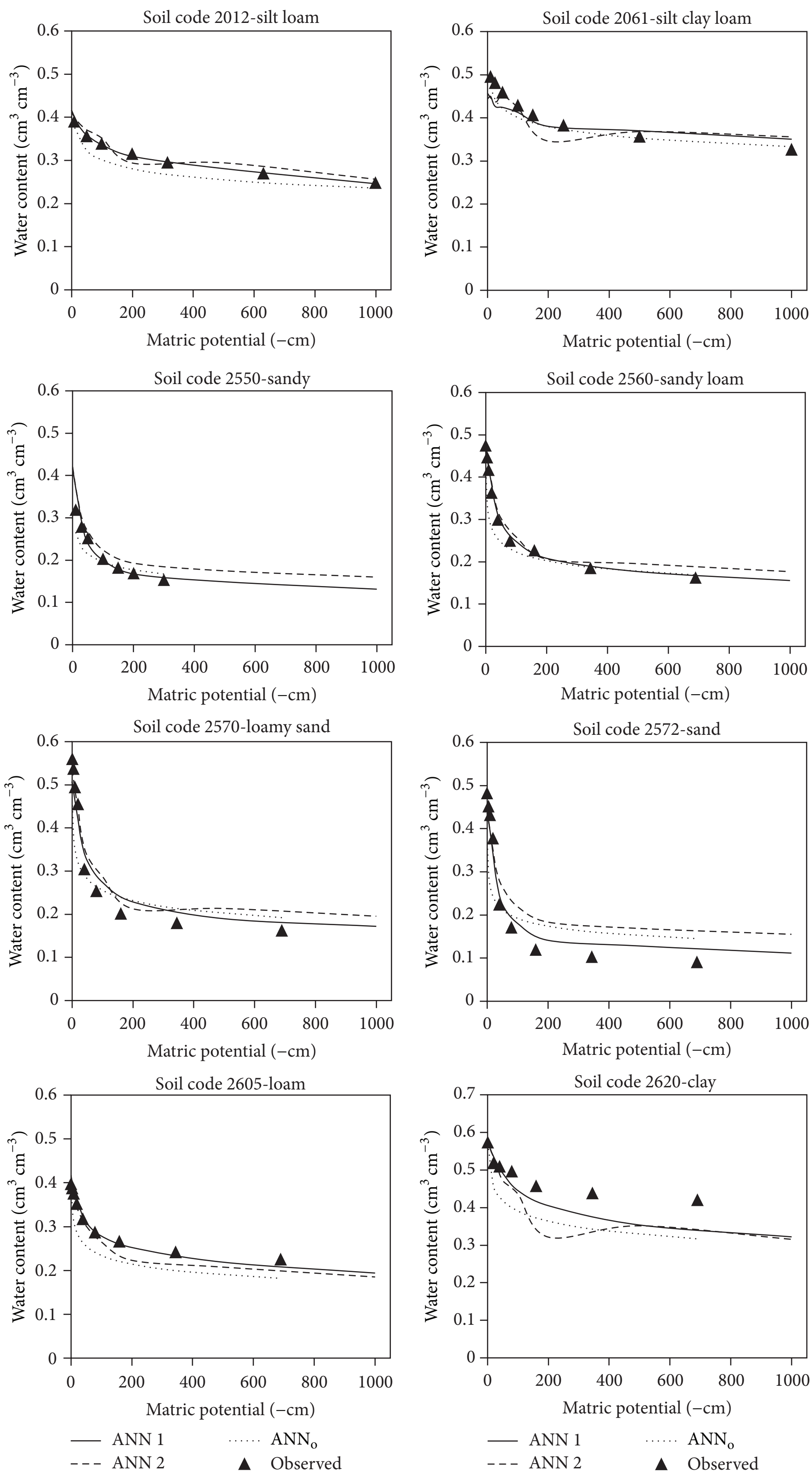

FIGURE 8: Observed and predicted soil water retention curves given by ANN models developed in this study.

Table 3: Sensitivity analysis of developed ANNs by substituting step by step each input parameter by its mean value.

ANN 1

h (−cm)	Model's RMSE	Bulk density		Particle density		Porosity		Sand		Silt		Clay	
		RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)
0	0.064	0.089	39.478	0.065	1.242	0.075	16.638	0.064	0.052	0.064	−0.523	0.069	7.5083
10	0.057	0.084	46.749	0.058	1.278	0.065	13.704	0.062	8.581	0.055	−3.819	0.058	1.718
20	0.056	0.079	40.550	0.056	0.859	0.065	16.952	0.067	12.082	0.056	−0.784	0.059	5.105
30	0.051	0.068	32.408	0.051	0.038	0.061	17.870	0.064	25.153	0.051	−0.019	0.056	8.416
50	0.049	0.064	27.561	0.050	−0.012	0.054	8.511	0.069	37.830	0.050	0.168	0.055	11.216
100	0.069	0.077	11.004	0.069	0.082	0.072	3.507	0.087	25.205	0.070	1.290	0.075	7.759
200	0.082	0.088	8.418	0.081	−0.690	0.084	2.557	0.100	23.078	0.081	−0.197	0.086	5.539
500	0.083	0.089	7.373	0.082	−0.605	0.084	1.654	0.096	15.955	0.083	0.320	0.088	6.276
1000	0.078	0.085	8.0672	0.077	−1.006	0.079	1.377	0.091	15.824	0.077	−1.178	0.083	6.041

ANN 2

h (−cm)	Model's RMSE	Bulk density		Particle density		Porosity		Sand		Silt		Clay	
		RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)
0	0.062	0.090	44.795	0.065	3.495	0.074	18.984	0.061	−1.4383	0.063	0.735	0.071	13.825
10	0.054	0.079	45.688	0.055	1.939	0.066	21.240	0.054	−0.346	0.061	11.0896	0.066	21.544
20	0.055	0.082	48.997	0.056	1.218	0.062	13.777	0.064	16.515	0.053	−4.713	0.056	1.897
30	0.055	0.079	44.116	0.055	0.319	0.060	8.939	0.065	18.440	0.055	1.269	0.057	4.422
50	0.060	0.075	25.927	0.061	1.625	0.065	9.170	0.068	14.038	0.063	6.347	0.067	11.943
100	0.075	0.086	13.880	0.075	0.528	0.077	2.454	0.082	8.963	0.080	6.047	0.082	8.753
200	0.098	0.103	5.356	0.098	0.001	0.102	3.909	0.107	9.005	0.101	3.209	0.105	7.686
500	0.087	0.095	8.460	0.087	−0.027	0.089	1.565	0.097	11.781	0.088	0.728	0.091	4.904
1000	0.084	0.086	3.339	0.083	−0.141	0.088	5.358	0.087	4.246	0.086	2.633	0.092	10.162

ANN_o

Model's RMSE	Bulk density		Particle density		Porosity		Sand		Silt		Clay		ln(h)	
	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)	RMSE	Var (%)
0.080	0.094	17.065	0.080	−0.075	0.085	5.7464	0.081	1.255	0.080	0.037	0.082	2.641	0.118	46.825

* Var (%) means the increasing (+) or decreasing (−) in the RMSE value of an output (matric potentials, h) due to the substitution of a specific input parameter by its mean value.

the range of intermediate matric potentials seems to be highly and significantly dependent of only sand percentages, followed by clay more evident in ANN 2. In our analysis, we considered a RMSE variation higher than 5% to be significant. For drier conditions, sand and clay are the inputs that most affected the predictions. This fact confirms what was previously mentioned that when flow in macropores dominates the processes (wet regime), soil structure is more relevant, while texture becomes important during micropores flow. Still, as already expected, silt did not cause significant changes in models performances for any range of matric potentials or ANN model, as well as particle density, probably due to its almost uniform distribution.

Analysing the ANN_o model, which included one more input parameter and has a single output, only a general inference on model performances could be done, since it is not possible to distinguish between distinct ranges of matric potentials. Thus, for this model, the most important input parameter was the natural logarithm of matric potentials, causing a decrease in model performances similar to those caused by bulk density in ANN 1 and ANN 2. Besides such an input, bulk density is the only other parameter that highly influenced ANN_o performance.

The objective of the sensitivity analysis was simply to prove which of the input parameters are more relevant and then exclude those which are not. For example, most studies presented in this study applied only bulk density in their PTFs. But we can cite the work of Bayat et al. [5] who used bulk density and total porosity in their PTFs and the work of Zacharias and Wessolek [7] who added the organic matter content to their PTF and proved that this parameter does not contribute to model's performance and that there is no need for using it to determine the soil water retention curve (although this parameter is extensively used in the literature with this purpose). It was expected that matric potential would be the most important parameter of the model of Haghverdi et al. [4], since it is well known that water content is directly influenced by the soil pressure head. But even in their model, bulk density was significant when compared to the other input parameters. One may note that is not possible to identify the range of matric potential for which the bulk density is important in their model, whereas, in ANN 1 and ANN 2, we could detect that bulk density is not relevant in drier conditions. The main point is that adding the matric potential as input parameter, the ANN_o model provided equivalent performance to our models, which need no information about the pressure head (we may not forget that the relation between water content and matric potential is not easy to be determined in the field or in laboratory). In summary, based on the sensitivity analysis, we may conclude that a new ANN model can be constructed only with sand, silt, clay, and bulk density, even though it may give satisfactory performance comparable to the model trained with measured data.

The results deserve a closer and more critical analysis of how many input parameters are really necessary for developing PTFs. For instance, the best RMSE values published in the researched literature presented in this study were found by Jain et al. [10], whose ANN was constructed solely by one input parameter, the measured matric potentials, and one output, the water contents. Other input combinations consisted of including previous points of suctions of the retention curve. This PTF was able not only to accurately predict the retention curve but also to decide which branch of the hysteresis to follow. On the other hand, the ANN_o model is believed to generate better performances because it dispenses any fitting and is applied only with measured points, which is a worthy question to verify.

4. Conclusions

Two different ANN models were developed for estimating the soil water retention curve using fitted data in contrast to measured data. This study was an attempt to investigate if employing fitted data from observed values was suitable for training ANNs. It was found that the ANN with fitted data and one output neuron for each matric potential, placed together in the same ANN structure, was the most accurate model and provided performances comparable to those presented in the literature. The results were even better than those obtained by the model trained with observed data, the pseudocontinuous method.

It is important to mention that this work did not aim to compare ANN structures and decide which one was the best among them. Instead, it aimed to show that fitting data could also be a satisfactory option to deal with small datasets beside the approach proposed by Haghverdi et al. [4]. Yet, it was assumed that fitting a water retention model to the measured data takes out some of the variability of the phenomenon, including measurement errors. Thus, it is possible that the uncertainties added by the van Genuchten model could be lower than measurement errors that are being corrected by fitting. That is the reason why ANN models trained with fitted data provided similar or even better performances than the pseudocontinuous method, which was trained with measured data.

ANNs trained with fitted data could be an alternative to parametric PTF (which also firstly needs to fit the parameters of a predefined model) and still find satisfactory predictions. An important drawback of parametric PTF models is that eventually some fitted parameters should be set constant or zero, as it is the case of the residual water content. This difficulty makes estimates with parametric PTFs generally less accurate than point PTFs. Additionally, better performances of the developed models could possibly be achieved if a larger dataset was used. As a suggestion for future researches, fitted data for training ANNs could be tested and compared to parametric and point ANNs and applied with a huge database. The tested ANNs in the present study tried to make it possible to develop PTF models incorporating whole available datasets, instead of considering only retention curve data that were obtained for the same chosen matric potentials.

In summary, this paper demonstrated not only that the neural networks trained with fitted data provided good predictions of the soil water retention curve but also that these models permit the identification of the most important input parameters, which was revealed to be not possible to be done with the pseudocontinuous method.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors are very thankful to CNPq Council for sponsoring this research and acknowledge Laura Bocco and all anonymous reviewers for their suggestions and improvements in this paper.

References

[1] M. Mukhlisin, A. El-Shafie, and M. R. Taha, "Regularized versus non-regularized neural network model for prediction of saturated soil-water content on weathered granite soil formation," *Neural Computing and Applications*, vol. 21, no. 3, pp. 543–553, 2012.

[2] R. B. Jana, B. P. Mohanty, and E. P. Springer, "Multiscale pedotransfer functions for soil water retention," *Vadose Zone Journal*, vol. 6, no. 4, pp. 868–878, 2007.

[3] H. Puhlmann and K. von Wilpert, "Pedotransfer functions for water retention and unsaturated hydraulic conductivity of forest soils," *Journal of Plant Nutrition and Soil Science*, vol. 175, no. 2, pp. 221–235, 2012.

[4] A. Haghverdi, W. M. Cornelis, and B. Ghahraman, "A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data," *Journal of Hydrology*, vol. 442-443, pp. 46–54, 2012.

[5] H. Bayat, M. R. Neyshabouri, K. Mohammadi, and N. Nariman-Zadeh, "Estimating water retention with pedotransfer functions using multi-objective group method of data handling and ANNs," *Pedosphere*, vol. 21, no. 1, pp. 107–114, 2011.

[6] C. D. Børgesen, B. V. Iversen, O. H. Jacobsen, and M. G. Schaap, "Pedotransfer functions estimating soil hydraulic properties using different soil parameters," *Hydrological Processes*, vol. 22, no. 11, pp. 1630–1639, 2008.

[7] S. Zacharias and G. Wessolek, "Excluding organic matter content from pedotransfer predictors of soil water retention," *Soil Science Society of America Journal*, vol. 71, no. 1, pp. 43–50, 2007.

[8] H. Merdun, Ö. Çinar, R. Meral, and M. Apan, "Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity," *Soil and Tillage Research*, vol. 90, no. 1-2, pp. 108–116, 2006.

[9] C. D. Børgesen and M. G. Schaap, "Point and parameter pedotransfer functions for water retention predictions for Danish soils," *Geoderma*, vol. 127, no. 1-2, pp. 154–167, 2005.

[10] S. K. Jain, V. P. Singh, and M. T. van Genuchten, "Analysis of soil water retention data using artificial neural networks," *Journal of Hydrologic Engineering*, vol. 9, no. 5, pp. 415–420, 2004.

[11] W. J. Rawls, Y. Pachepsky, and M. H. Shen, "Testing soil water retention estimation with the MUUF pedotransfer model using data from the southern United States," *Journal of Hydrology*, vol. 251, no. 3-4, pp. 177–185, 2001.

[12] J. H. M. Wösten, A. Lilly, A. Nemes, and C. Le Bas, "Development and use of a database of hydraulic properties of European soils," *Geoderma*, vol. 90, no. 3-4, pp. 169–185, 1999.

[13] H. Vereecken, J. Maes, J. Feyen, and P. Darius, "Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content," *Soil Science*, vol. 148, no. 6, pp. 389–403, 1989.

[14] C. Manyame, C. L. Morgan, J. L. Heilman, D. Fatondji, B. Gerard, and W. A. Payne, "Modeling hydraulic properties of sandy soils of Niger using pedotransfer functions," *Geoderma*, vol. 141, no. 3-4, pp. 407–415, 2007.

[15] R. H. Brooks and A. T. Corey, "Properties of porous media affecting fluid flow," *Journal of Irrigation and Drainage*, vol. 92, pp. 61–88, 1964.

[16] M. T. van Genuchten, "A closed-form equation for predicting the hydraulic conductivity of unsaturated soils," *Soil Science Society of America Journal*, vol. 44, no. 5, pp. 892–898, 1980.

[17] L. Baker and D. Ellison, "Optimisation of pedotransfer functions using an artificial neural network ensemble method," *Geoderma*, vol. 144, no. 1-2, pp. 212–224, 2008.

[18] F. J. Leij, W. J. Alves, M. T. van Genuchten, and J. R. Williams, *Unsaturated Soil Hydraulic Database, UNSODA 1.0 User's Manual. Report EPA/600/R-96/095*, US Environmental Protection Agency, Ada, Okla, USA, 1996.

[19] A. Nemes, M. G. Schaap, F. J. Leij, and J. H. M. Wösten, "Description of the unsaturated soil hydraulic database UNSODA version 2.0," *Journal of Hydrology*, vol. 251, no. 3-4, pp. 151–162, 2001.

[20] M. T. van Genuchten, F. Simunek, F. J. Leij, and M. Sejna, "The RETC code (version 6.02) for quantifying the hydraulic functions of unsaturated soils," 2009, http://www.hydrus3d.com.

[21] E. J. W. Koekkoek and H. Booltink, "Neural network models to predict soil water retention," *European Journal of Soil Science*, vol. 50, no. 3, pp. 489–495, 1999.

[22] P. Santra and B. S. Das, "Pedotransfer functions for soil hydraulic properties developed from a hilly watershed of Eastern India," *Geoderma*, vol. 146, no. 3-4, pp. 439–448, 2008.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[24] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in *IRE WESCON Convention Record*, part 4, pp. 96–104, IRE, New York, NY, USA, 1960.

[25] Y. Abbasi, B. Ghanbarian-alavijeh, A. M. Liaghat, and M. Shorafa, "Evaluation of pedotransfer functions for estimating soil water retention curve of saline and saline-alkali soils of Iran," *Pedosphere*, vol. 21, no. 2, pp. 230–237, 2011.

[26] M. Gevrey, I. Dimopoulos, and S. Lek, "Review and comparison of methods to study the contribution of variables in artificial neural network models," *Ecological Modelling*, vol. 160, no. 3, pp. 249–264, 2003.

[27] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," *Ecological Modelling*, vol. 178, no. 3-4, pp. 389–397, 2004.

[28] K.-H. Liao, S.-H. Xu, J.-C. Wu, S.-H. Ji, and Q. Lin, "Assessing soil water retention characteristics and their spatial variability using pedotransfer functions," *Pedosphere*, vol. 21, no. 4, pp. 413–422, 2011.

[29] A. Mermoud and D. Xu, "Comparative analysis of three methods to generate soil hydraulic functions," *Soil & Tillage Research*, vol. 87, no. 1, pp. 89–100, 2006.

[30] N. Romano and A. Santini, "Effectiveness of using pedotransfer functions to quantify the spatial variability of soil water retention characteristics," *Journal of Hydrology*, vol. 202, no. 1–4, pp. 137–157, 1997.