Prediction of water retention of soils from the humid tropics by the non-parametric k-nearest neighbor approach

Yves-Dady Botula[a,b,*], Attila Nemes[c], Paul Mafuka[d], Eric Van Ranst[b], Wim M. Cornelis[a]

Abstract

Non-parametric approaches such as the k-Nearest Neighbor (k-NN) approach are nowadays considered as attractive tools for pedotransfer modeling in hydrology. However, non-parametric approaches have not been applied so far to predict water retention of highly weathered soils in the humid tropics. Therefore, the objectives of this study are: to apply the k-Nearest Neighbor (k-NN) approach to predict soil water retention in a humid tropical region; to test its ability to predict soil water content at eight different matric potentials; to test the benefit of using more input attributes than most previous studies did and their combinations; to discuss the importance of particular input attributes in the prediction of soil water retention at low, intermediate and high matric potentials and to compare this approach to two published tropical pedotransfer functions (PTFs) based on multiple linear regression (MLR). The overall estimation error ranges generated by the k-NN approach were statistically different but comparable to the two examined MLR PTFs. When the best combination of input variables (i.e. sand+silt+clay+bulk density+cation exchange capacity) is used, the overall error is remarkably low: 0.0360 to 0.0390 m^3 m^{-3} at the dry and the very wet ranges, and 0.0490

24 to 0.0510 m^3 m^{-3} at the intermediate range (i.e. -3 to -50 kPa) of the soil water retention
25 curve. This k-NN variant can be considered as a competitive alternative to more classical
26 equation-based PTFs due to the accuracy of the water retention estimation and, as added
27 benefit, its flexibility to incorporate new data without the need to redevelop new
28 equations. This is highly beneficial in developing countries where soil databases for
29 agricultural planning are at present sparse, though slowly developing.

30

1. Introduction

The unsaturated soil hydraulic functions are important parameters in many pedological, hydrological, ecological and agricultural studies (Rajkai et al., 2004). However, direct measurements of such parameters are still expensive and time-consuming especially for studies at a regional scale (Vereecken, 1995; Pachepsky et al., 2006; Guber et al. 2006). Medina et al. (2002) stated that in developing countries, there are additional problems associated with this task, ranging from personnel training to acquisition of the necessary equipment. Therefore, an attractive alternative to the direct and often cumbersome measurements of soil hydraulic properties is their estimation by so-called pedotransfer functions (PTFs). Bouma (1989) described the term pedotransfer function (PTF) as "translating data we have into what we need". PTFs thus relate more easily measurable soil data and/or other data routinely measured or registered in soil surveys with hydraulic parameters in a statistical sense (Bouma and van Lanen, 1987; Bouma, 1989; van den Berg et al., 1997).

Another alternative to obtain estimates or approximates of hydraulic properties is inverse modeling. Inverse procedures have the potential to yield information about soil hydraulic conductivity and water retention over a wide range of matric potentials from a single infiltration experiment (Schwartz and Evett, 2002). Briefly, the multistep outflow method applies inverse modeling technique for indirect estimation of both water retention and hydraulic conductivity curves in a single transient drainage experiment (van Dam et al., 1994). The soil hydraulic parameters of an analytical function for the soil water retention curve (SWRC) (e.g. van Genuchten, 1980) or for hydraulic conductivity (e.g. Mualem,

1976) are determined by matching experimental observations of transient water flow with numerical modeling results. In simple words, the estimated parameters are the solutions of an inverse problem. The latter results in determining causes that are unknown a priori, based on observations of their effects. Hopmans et al. (2002) presented a comprehensive review of inverse modeling for estimation of soil hydraulic properties, including one-step and multistep methods. While this technique can yield rather accurate set of effective soil hydraulic properties, its feasibility is limited for large scale applications and/or when intended to be used in areas or countries with scarce resources.

When applying pedotransfer modeling or inverse modeling to obtain estimates or approximates of hydraulic properties, we should bear in mind that soils from tropical regions are vastly different from soils from temperate regions (e.g. van den Berg et al., 1997; Hodnett and Tomasella, 2002; Minasny and Hartemink, 2011; Botula et al., 2012). Botula et al. (2012) evaluated the ability of some selected PTFs to predict θ_{33kPa} and $\theta_{1500kPa}$ of a limited dataset of soils from the Lower Congo, the south-western part of the Democratic Republic of Congo (D.R. Congo) located in the humid tropics. They found that the temperate-climate PTFs of Gupta and Larson (1979) and Rawls and Brakensiek (1982) largely overestimated water retention of soils in the Lower Congo. These PTFs were derived based on temperate-climate soils from across the USA. On the other hand, they demonstrated that the tropical-climate PTFs of Hodnett and Tomasella (2002) performed well compared to aforementioned temperate-climate PTFs. Hodnett and Tomasella (2002) used a part of the IGBP-DIS soil database obtained from ISRIC-World Soil Information in Wageningen (the Netherlands) to derive PTFs for predicting the four

parameters of the van Genuchten (1980) equation. The authors referred to this development dataset as the IGBP/T dataset which exclusively contained soils from tropical climates. Botula et al. (2012) attributed the poor predictive performance of the "temperate" PTFs to the differences in soil properties and mineralogy between the test dataset and the dataset used to develop these PTFs. They recommended that more efforts should be done to develop specific PTFs to predict water retention of soils in the tropics. Schaap (2005) wrote that "with the exception of a few studies, hydraulic data and corresponding indirect methods about tropical soils are a virtual *terra incognita*". This situation has not changed much by today. Also Minasny and Hartemink (2011) noted that limited efforts are devoted to the prediction of properties of soils in the tropics where the need for accurate and up-to-date soil property information is even more urgent than elsewhere. They identified various soil properties used to predict the soil water retention curve (SWRC) in the tropics such as sand, silt, clay, bulk density (BD), organic carbon/matter (OC/OM), pH, cation exchange capacity (CEC), dithionite-citrate-bicarbonate, extractable iron (DCB-Fe) and aluminum (DCB-Al), but finally selected soil texture, BD and OC to develop PTFs to predict water content at -10, -33 and -1500 kPa. The development dataset and the validation dataset exclusively contained soils from the tropics. These soil datasets are also part of the international IGBP-DIS soil database obtained from ISRIC.

Despite the limited efforts in data collection and harmonization for soils from the humid tropics (where most of the developing countries are located) compared to temperate areas, large tropical soil databases will steadily grow. With the emergence of such large

databases, classical statistical methods such as multiple linear regressions (MLR) may show limitations as important trends may not be detected, whereas others may falsely be given much emphasis. Therefore, there is a need to promote data-mining or pattern-recognition techniques which are flexible enough to handle huge amounts of data and detect important trends which may be hidden to classical statistical methods such as MLR.

Even though classic PTFs based on the MLR approach have been widely used to predict water retention in the past, PTFs based on pattern-recognition approaches have gained popularity. This is particularly because they present the advantage of including new soil information without the constraint of redeveloping new equations to fit the new soil dataset. This flexibility in incorporating new soil data is highly beneficial in tropical regions particularly for developing countries, where continuously developing soil databases are highly demanded for pedological, agricultural and ecological studies. Pattern-recognition techniques belong to the group of data-driven, data-mining or machine-learning techniques, in contrast with MLR which is based on predefined mathematical functions. Recently, three pattern-recognition techniques have been used with success in studies related to unsaturated soil hydrology: Artificial Neural Networks (ANN), Support Vector Machines (SVM) and the k-Nearest Neighbor (k-NN) technique. Mucherino et al. (2009) provided an elaborated review of these data-mining techniques and on their application in various agriculture- and environment-related fields. For further information on the ANN and SVM techniques, we refer the reader to Hecht-Nielsen (1990), Haykin (1994), Vapnik (1995, 1998) and Noble (2006).

In this study, we use the k-NN technique which is considered as one of the most attractive pattern-recognition algorithms by several authors (e.g. Buishand and Brandsma, 2001; Bannayan and Hoogenboom, 2009). It is referred to as a "lazy learning algorithm" because it passively stores the data until the time of application. All calculations are performed "real-time" i.e. only when estimations need to be generated. Application of the k-NN technique means identifying and retrieving the most similar instances to the target object from the multi-dimensional feature (input variable) space of the set of stored instances, and classifying the target object based on similarities in their input attributes and using a pre-defined weighting scheme. More theoretical details on this similarity-based approach are given in Dasarathy (1991).

Nemes et al. (1999) used a k-NN variant – which they termed the "similarity technique" to estimate missing soil particle size distribution (PSD) points from other existing PSD points in order to harmonize data of the European HYPRES database (Wösten et al. 1999). Jagtap et al. (2004) used a k-NN technique to estimate the drained upper limit and lower limit of plant water availability from soil water retention data measured in-situ. Nemes et al. (2006a) provided several examples of applications of the k-NN techniques in hydrologic simulation and developed another variant of the k-NN technique to estimate soil water retention at two matric potentials. They also performed a detailed sensitivity analysis of this technique (Nemes et al., 2006b). The newly developed k-NN algorithm proved its robustness in different scenarios. Based on the satisfactory results yielded by their k-NN algorithm, Nemes et al. (2008) developed a user-friendly software called "k-

Nearest" to estimate θ_{-33kPa} and $\theta_{-1500kPa}$ with the option of estimating the uncertainty of the prediction using data re-sampling. Elshorbagy et al. (2010a,b) conducted a detailed study of the predictive capabilities of data-driven modeling techniques in hydrology, and identified the k-NN technique as an attractive modeling technique for hydrological applications because of its high level of flexibility, due to reasons mentioned above. Nemes et al. (2006a) specifically refer to the k-NN method working with patterns of similarities instead of fitting equations to data, and its real-time application giving users the flexibility to alter the underlying data or the calculation scheme. Gharahi Ghehi et al. (2012) recently applied the k-NN approach for predicting bulk density of Rwandese soils in the humid tropics.

When predicting hydraulic properties on the basis of existing databases for training by data-driven models, Perkins and Nimmo (2009) stressed the necessity of high quality databases. They indicated that an obvious problem occurs when the available database has few or no data for samples that are closely related to the region of interest. This is classically the case when a dataset of soils from temperate areas is used as a training dataset to predict hydraulic properties of soils from tropical regions. In their sensitivity analysis, Nemes et al. (2006b) used separate datasets from the USA, Europe and Brazil and found that when using a dataset of "temperate soils" as a training dataset to predict water retention of "tropical" soils from Brazil, estimations were significantly worse than for other examined dataset pairs, with bias errors amounting to an undesirable 0.10 m^3 m^{-3}. As point of future research, Nemes et al. (2006a) recommended testing the ability of the k-NN approach to predict soil water retention based on datasets from different regions

of the world, but an application that uses an international collection of soils from the humid tropics is still lacking.

Point estimation PTFs are usually limited to estimating only a few points on the water retention curve, most frequently two or three points. Among such applications are estimations using k-NN. In their application, Nemes et al. (2006a) predicted water content by their k-NN variant at -33 kPa and -1500 kPa matric potentials, using a small number of input attributes: texture (Sand+Silt+Clay, designated here as SSC), OM and BD. Recently, Patil et al. (2012) used the k-NN software developed by Nemes et al. (2008) to estimate θ_{-33kPa} and $\theta_{-1500kPa}$ of 157 swelling-shrinking soils in India in order to derive their available water capacity. These matric potentials were also used by numerous other studies (e.g. Givi et al., 2004; Reichert et al., 2009; Minasny and Hartemink, 2011; Botula et al., 2012). The rationale is that these two points are meant to be used as approximates to water retention at *field capacity* (FC) (θ_{-33kPa}) and *permanent wilting point* (PWP) ($\theta_{-1500kPa}$), in order to calculate available water holding capacity or to parameterize bucket-type agronomic or water balance models. This raises two considerations that were of significance when initiating this study.

First, it is still debated what, if any, matric potential is a good representation of conditions at/near field capacity. It appears to be generally affected by a number of factors, among them soil texture. Apart from field experiments (Ottoni Filho and Ottoni, 2010), and data mining studies (Nemes et al. 2011), Twarakavi et al. (2009) also demonstrated this dilemma using inverse modeling. However for tropical soils, several authors (e.g. Sharma

192 and Uehara, 1968; Pidgeon, 1972; Babalola, 1979; Lal, 1978; Reichardt, 1988) suggested
193 that water content at -10 kPa represents FC better than water content at -33 kPa which is
194 more frequently adopted by authors working with soils of temperate climate. The soil-
195 water relation of well-aggregated kaolinitic soils under tropical climate can be markedly
196 different from that in soils with permanent charge minerals in temperate regions. Heavy-
197 textured soils dominated by kaolinite and sesquioxides have SWRCs which in some
198 respects resemble those of sandy soils (Sharma and Uehara, 1968), although they show
199 higher porosity. In aggregated highly weathered soils (e.g. Ferralsols), water can reside in
200 large inter-aggregate pores and fine intra-aggregate pores. Under gravitational forces,
201 water in the large pores move rapidly and FC is attained at high matric potentials,
202 generally between -10 kPa and -15 kPa. Field capacity is attained at this high matric
203 potential because the hydraulic conductivity at this potential is very low, much like that
204 of a sandy soil. It may therefore be advisable to have information on water content at
205 higher matric potentials than -33 kPa, when it comes to supporting studies in the humid
206 tropics that concern the unsaturated zone. At the same time, according to the studies cited
207 above, water content at -1500 kPa can still be considered as an approximation of the
208 permanent (PWP).
209
210 The second consideration is that when two or three points are estimated on the SWRC, it
211 allows no or only limited (constrained) use of popular water retention models like the
212 models of van Genuchten (1980) or Brooks and Corey (1964). Pedotransfer functions that
213 estimate parameters of such models offer a solution to this dilemma; however, it was
214 found by Tomasella et al. (2003) that estimating SWRC points followed by curve fitting

yielded more accurate results than estimating curve parameters and reading water content values at particular matric potentials off the fitted curve. Hence, we have chosen to estimate a number of water retention points that will facilitate the subsequent use of both point and parameterized SWRC. Overall, to be able to fit a complete SWRC, six to eight measured or estimated water retention points are recommended as the SWRC models more commonly used (e.g. Brooks and Corey, 1964; van Genuchten, 1980) have four or more fitting parameters (Tomasella et al., 2000; Cornelis et al., 2005). Until now, no study has been published that estimates water content at more than two matric potentials using the k-NN method. It was facilitated by the databases available for this study that we estimate up to eight SWRC points.

Therefore, the objectives of this paper are: (1) to apply a non-parametric approach to obtain estimations of water content of soils for a tropical region, based on an international database of soils from the humid tropics and using an adaptation of the k-NN algorithm developed by Nemes et al. (2006a), (2) to test the ability of the k-NN algorithm to predict several points of the SWRC (i.e. water content at eight different matric potentials) from the wet to the dry range simultaneously, (3) to use a range of input attributes and determine the influence of several combinations of input attributes on the ability of the k-NN approach to predict water content at those matric potentials, (4) to discuss the importance of particular input attributes in the estimation of soil water content at low, intermediate and high matric potentials and (5) to compare the prediction performance of the proposed k-NN variant and two aforementioned MLR PTFs which were developed using datasets from the tropics, similarly extracted from the international IGBP database.

2. Materials and Methods

2.1. Soil datasets

In this study, a dataset of 534 soils from tropical regions was used as the reference/training dataset for the k-NN estimations. These soil samples are part of the IGBP-DIS international database from ISRIC (Tempel et al., 1996). By tropical regions, we mean the regions situated between 25°N and 25°S and mainly under the (sub)-humid climates. Soils within the tropics but in temperate climates due to altitude or in dry areas are not included in the selected dataset. This "tropical" dataset is referred to here as the IGBP-Trop dataset. It contains highly weathered soils such as Ferralsols (20.4%), Acrisols (11.6%) and Nitisols (4.7%), and other soils like Cambisols (14.2%), Andosols (6.4%), Luvisols (6%), Gleysols (4.7%), Phaeozems (4.3%), Fluvisols (2.8%), Vertisols (2.6%), Arenosols (2.4%) among others (IUSS Working Group WRB, 2006). Undisturbed and disturbed soil samples were collected under different land uses and under various depths.

The associated digital database contains, among other attributes, water content data at eight different matric potentials (0, -1, -3, -10, -20, -50, -250 and -1500 kPa). Tempel et al. (1996) provided the necessary references concerning the different analytical methods used to derive the soil physical and chemical properties recorded in the database.

A dataset of 139 soils from the Lower Congo, the south-western part of the D.R. Congo was used as an independent dataset to test the predictive ability of the k-NN approach. These soils are mainly highly weathered soils under the humid tropics classified as

Ferralsols, Acrisols and Nitisols (IUSS Working Group WRB, 2006) but other Soil Groups such as Umbrisols and Arenosols (IUSS Working Group WRB, 2006) were also represented. The 139 selected soil samples were not part of the IGBP-DIS database. Undisturbed soil samples were collected in 100 cm^3 Kopecky rings under different land uses (savannah, forest, agricultural fields and old quarries) and under various depths in the soil profile. For the undisturbed samples, the SWRC data pairs were determined from the wet to the dry range at eight different matric potentials: -1, -3, -6, -10, -20, -33, -100 and -1500 kPa. The hanging water-column method was used for matric potentials between -1 and -10 kPa using the sand box apparatus (Eijkelkamp Agrisearch Equipment, Giesbeek, the Netherlands), whereas for matric potentials between -20 and -1500 kPa, pressure chambers (Soil Moisture Equipment, Santa Barbara, CA) were used, following the procedures described in Cornelis et al. (2005). The coupled matric potential-water content pairs represent single measurements on single samples. Matric potentials at 0, -50 and -250 kPa used in the IGBP-Trop database were missing in the Lower Congo database. Therefore, they were derived by curve fitting as follows: (1) a continuous curve was fitted through the discrete set of measured (available) water retention points using the van Genuchten (1980) function, and (2) fitted values of water contents at the missing matric potentials (0, -50 and -250 kPa) were calculated from the resulting continuous equation. The physico-chemical characteristics of all soil samples (fine earth) were determined using standard methods described in detail by Van Ranst et al. (1999). During these analyses, PSD (by the pipette method of Köhn, 1929), OC, pH, and CEC were determined on the same soil samples that were previously used for SWRC measurements.

Soil properties selected for use in this study were the following: sand (50–2000 μm), silt (2–50 μm), and clay content (< 2 μm) according to the USDA classification system (USDA, 1951), BD, OC, pH, CEC and retained (volumetric) water content (θ) at eight different matric potentials: 0, -1, -3, -10, -20, -50, -250 and -1500 kPa. Any entries that showed obvious inconsistency in physical and/or hydraulic data (e.g. sand + silt + clay ≠ 1; $\{[1 - BD/2.65] - \theta_{0kPa}\} < 0$; $\theta_{xkPa} < \theta_{ykPa}$ when x kPa > y kPa) were excluded from the reference/training dataset and the test dataset. Figure 1 shows the textural distribution of the IGBP-Trop and the Lower Congo datasets.

2.2. k-Nearest Neighbor technique

The k-NN algorithm used in this study has been adapted from the variant developed by Nemes et al. (2006a). The same algorithm was used in this study but has been expanded to use more input and output attributes and the design parameters of the algorithm had been reevaluated for the current application. The implementation was done in the MATLAB R2010a environment (The MathWorks, Inc., Hill Drive Natick, MA).

2.2.1. Rationale

The k-NN technique does not use any predefined mathematical function to estimate a certain response attribute like classic MLR PTFs do. It does not appear to rely on any stringent assumptions about the underlying data, and can adapt to any situation (Hastie et al., 2009). The k-NN approach consists of finding the *k* number of nearest neighbors from a reference dataset to each soil in the test dataset in terms of their selected input attributes. The similarity distance to the target soil is measured in terms of Euclidean

distance after normalization and rescaling of the soil attributes data in the reference dataset following a specific procedure. This is done to assure that different input attributes will receive equal weight. In ascending order of their (normalized) similarity distance to the target soil, soils will be sorted in the reference dataset. The number of selected nearest soil instances (k) needs also to be optimized following a specific procedure. Once the nearest neighbors are identified and sorted, distance-dependant weights are assigned to them and the response attribute is formulated and outputted as the weighted average of the response attributes of the selected nearest neighbors. More methodological and calculation details on the whole procedure are given below.

2.2.2. Selection of the nearest neighbors to the target soil

An external training (reference) dataset containing information on a wide variety of soils is searched for soils (instances) that are most similar to the target soil, based on the selected input attributes or features. Similarity between the target soils and the known instances is measured in terms of a metric considered here as the Euclidean distance:

$$d_i = \sqrt{\sum_{j=1}^{x} \Delta a_{ij}^2} \qquad [1]$$

where d_i is the "distance" of the i^{th} soil from the target soil, and Δa_{ij} is the difference of the i^{th} soil from the target soil in the j^{th} soil attribute.

In ascending order of their distance to the target soil, soils of the reference dataset will be sorted.

2.2.3. Normalization of soil data

Soils present some properties (attributes) which differ in their order of magnitude and/or range. For instance, a non-organic soil can have 100% of sand but should not have more than 18% of OC (Soil Survey Staff, 1975). Therefore, a unit difference in OC is expected to be more significant than the same unit difference in sand content. Therefore, a normalization procedure was applied on the soil properties data before they were used to calculate the Euclidean distance given in Eq. [1]. Normalizing the soil attributes has the benefit of lowering bias toward one soil attribute or the other. All input attributes were first transformed to temporary variables $a_{ij(temp)}$ with a distribution having zero mean and standard deviation of 1 by the following classic formula:

$$a_{ij(temp)} = \left(\left(a_{ij}\right) - \overline{a_j}\right) / \sigma\left(a_j\right) \quad [2]$$

where a_{ij} is the value of the j^{th} attribute of the i^{th} soil, and $\bar{a}_j$ and $\sigma(a_j)$ are the mean and standard deviation of the observed values of the j^{th} attribute in the reference dataset.

Secondly, the difference between the minimum and maximum of the aforementioned temporary variables was then examined in order to identify the soil attribute that shows the widest range of transformed (temporary) values. This allows a scaling of the temporary variables to obtain zero mean and the same minimum-maximum range in the data of all attributes:

$$a_{ij(trans)} = a_{ij(temp)} \left(Max\left\{range\left[a_{j=1(temp)}\right], \ldots, range\left[a_{j=x(temp)}\right]\right\}\right) / range\left[a_{j(temp)}\right] \quad [3]$$

where $a_{j(temp)}$ is the data of the j^{th} soil attribute normalized using Eq. [2], and $a_{ij(trans)}$ is the final transformed value of the j^{th} attribute of the i^{th} soil. Eventually, $a_{ij(trans)}$ values derived from Eq. [3] were used as input in our k-NN algorithm.

354

355 2.2.4. Application of a distance-dependent weighing system

356 A weighing procedure that accounts for the distribution of the distances of the selected *k*
357 neighbors from the target soil was applied. Weights of each selected neighbor were
358 computed as:

359

360 $$w_i = d_{i(rel)} / \sum_{i=1}^{k} d_{i(rel)} \quad [4]$$

361

362 where *k* is the number of neighbors selected, w_i is the weight associated to the i^{th} nearest
363 neighbor, and $d_{i(rel)}$ is the relative distance of the i^{th} selected neighbor calculated as:

364

365 $$d_{i(rel)} = \left(\sum_{i=1}^{k} d_i / d_i \right)^p \quad [5]$$

366

367 where d_i is the distance of the i^{th} selected neighbor computed using Eq. [1], and *p* is a
368 power term to account for different possible weight/distance relationships.
369 Therefore, the predicted water retention at a given matric potential corresponds to the
370 (distance-dependent) weighted sum of observed water retention values of the selected
371 nearest neighbors.

372

373 **2.3.** Design parameters *k* and *p* for the k-NN algorithm

374 There are two design-parameters of the k-NN algorithm that were used, namely the *k* and
375 the *p* terms. The *k* term refers to the number of similar soils to be selected from the

reference dataset to estimate the output attributes for each target soil, while the p term determines the weight-distance relationship that determines the contribution of each of the k reference samples to the estimation of the output attribute, depending on their degree of similarity to the target soil.

Nemes et al. (2006a) indicated that the best combination of k and p values i.e. the one leading to the lowest overall prediction error (expressed by the root mean square difference, RMSD detailed in Eq. [9]) should be selected and that such a choice may depend on the size of the reference dataset. They tested this assumption on different dataset sizes, i.e. N_r=100, 200, 400 and 800 and derived two different functions for k and p which are dependent of the size N_r of the reference dataset:

$$k = 0.655 N_r^{0.493} \quad [6]$$

$$p = 0.767 N_r^{0.049} \quad [7]$$

However, they warned that the relationship between N_r, k and p in Eq. [6] and Eq. [7] were set empirically and may not be optimal for other datasets. They recommended testing the settings of the k and p parameters for particular applications.

In this study, we re-optimized the two parameters using an approach similar to the one used by Nemes et al. (2006a). We determined what influence, if any, different k and p values have on the prediction performance of the k-NN algorithm in a tropical context i.e. when soils from the Lower Congo are used as test dataset and the international IGBP-

399 Trop dataset as training dataset. To avoid possible bias towards one or another set of
400 inputs, all pre-determined input variables (i.e. SSC+BD+OC+pH+CEC) to estimate all
401 the eight water retention points as outputs were considered. Then, all the corresponding
402 RMSDs were computed and plotted for a visual examination and the best combination of
403 *k* and *p* values was selected for this particular application. As the difference in RMSDs
404 between two subsequent *p* values is rather small, we decided to consider a change of p
405 from 0.5 to 2.5, with increments of 0.5, whereas the values of *k* were changed from 0 to
406 50, with increments of 1. The optimized combination of *k* and *p* was then used in further
407 calculations.
408
409 **2.4.** Ensemble of k-NN estimations
410 We experimented with the influence of the reference dataset size, similarly to Nemes et
411 al. (2006a), and so samples were drawn to be included in the development/reference
412 datasets of 100, 200, 300, 400 and 534 samples (i.e. all samples with available data). All
413 random data selections were repeated 100 times to allow the development of an ensemble
414 of water retention estimations. For each dataset size, the development/reference dataset
415 was randomly sampled 100 times at 80% resampling rate i.e. a different subsample
416 representing 80% of the development/reference dataset was used in each of the 100
417 replicates.
418
419 An ensemble of estimations has numerous advantages: the impact of any single replicate
420 (i.e., any particular dataset division) on the final estimation results can be minimized
421 when a sufficiently large number of replicates are used. Moreover, generation of an

ensemble of estimations allows the quantification of the uncertainty of estimates which can be used in statistical analyses and/or be inputted in simulation models. Quantification of uncertainty in estimates of soil hydraulic properties by PTFs and its effects in various simulation models has been studied by several authors (Finke et al., 1996; Nemes et al., 2003; Deng et al., 2009; Loosvelt et al., 2011; Moeys et al., 2012) who indicated that the uncertainty associated with hydraulic PTFs should be taken into account when evaluating simulation results yielded by a given model.

In this study, we found empirically that 100 replicates are sufficient to make the effect of any single replicate on the estimations negligible. Therefore, in this study we used 100 replicates in the algorithm and any statistical measures were computed based on those 100 replicates. However, we also examined the minimum (optimized) number of replicates for each of the different dataset sizes (N_r = 100, 200, 300, 400 and 534).

2.5. Input and output attributes used

In this paper, we have selected a wide range of soil attributes as potential predictors. These soil properties are not only used by several authors for the determination of "tropical" PTFs but are also important to characterize soils in the (sub)-humid tropics: sand, silt, clay, BD, OC, pH, CEC. Fourteen different combinations of these input attributes were considered to generate estimations in a hierarchical structure, in order to evaluate which, if any, of the variable combinations will yield systematically better estimates. The output attributes are water content at eight different matric potentials, namely at 0, -1, -3, -10, -20, -50, -250 and -1500 kPa. This means that we estimate more

445 water retention points simultaneously, in the wet, the intermediate and the dry range of
446 the SWRC.
447
448 **2.6.** Evaluation criteria
449 Three statistical measures were selected to assess the predictive ability of the k-NN
450 algorithm at a given matric potential: the mean difference (MD), the root mean square
451 difference (RMSD) and the coefficient of determination (R^2):
452
453 $$\mathrm{MD} = \frac{1}{N_t}\sum_{i=1}^{N_t}(\theta_{p_i} - \theta_{m_i}) \quad [8]$$
454
455 $$\mathrm{RMSD} = \sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}(\theta_{p_i} - \theta_{m_i})^2} \quad [9]$$
456
457 $$\mathrm{R}^2 = \frac{\left(\sum_{i=1}^{N_t}(\theta_{p_i} - \overline{\theta_{p_i}})(\theta_{m_i} - \overline{\theta_{m_i}})\right)^2}{\sum_{i=1}^{N_t}(\theta_{p_i} - \overline{\theta_{p_i}})^2(\theta_{m_i} - \overline{\theta_{m_i}})^2} \quad [10]$$

458 where θ_{pi} is the predicted volumetric water content for soil sample i (m^3 m^{-3}), θ_{mi} is the
459 measured volumetric water content for soil sample i (m^3 m^{-3}), and N_t is the number of
460 samples in the test dataset.
461
462 **2.7.** Comparison with two published "tropical" PTFs
463 The prediction performance of the proposed k-NN approach was compared with the
464 prediction performance of the MLR PTFs of Hodnett and Tomasella (2002) and Minasny

and Hartemink (2011) based on their RMSD values. As mentioned above, the PTFs of Hodnett and Tomasella (2002) predict the parameters of the van Genuchten (1980) equation based on basic soil properties (texture, BD, OC, pH and CEC). Therefore, they allow the calculation of water content at any given matric potential. On the other hand, the PTFs of Minasny and Hartemink (2011) predict water content from texture, BD and OC at three matric potentials: -10, -33 and -1500 kPa. In the present study, only results for -10 and -1500 kPa will be considered in the comparison with the k-NN approach as water content at -33 kPa is lacking in the IGBP-Trop dataset.

3. Results and Discussion

Box-plots of the selected soil attributes, for the reference/training dataset (IGBP-Trop) and for the test dataset (Lower Congo) are given in Fig. 2. Based on these soil attributes, it can be seen that both the reference and the test datasets contain data of a wide range of soils.

3.1. Ensembles of k-NN estimations

To find a minimum number of ensembles to obtain a stable RMSD based on the IGBP-Trop dataset, we plotted the running (cumulative) RMSD values against the total number of ensemble members after each replication dataset had been applied to make estimations. The magnitude and the evolution of the RMSD values with the number of ensembles M differ from one matric potential to the other but the difference seems to be marginal in practice (Fig. 3). It can be seen from Fig. 3 that using 30 ensemble members gives stable and satisfactory results using various proportions of the IGBP-Trop dataset as reference data. Using more than 30 replicates, we found practically no change for dataset size N_r=100, 200, 300, 400 and 534. The same observation was made in the wet, the intermediate as well as in the dry range of the SWRC.

Nemes et al. (2006b) determined that the sufficient minimum number of ensembles for the U.S. NRCS-SCS and the HYPRES datasets were 30 and 50 respectively. They found that using more than 30 or 50 ensembles respectively, the effect of adding more ensemble members did not yield any significant changes to the outcome of the estimations, regardless of the reference dataset size. Using the ANN technique, Parasuraman et al. (2006) found also that 30 ensemble members was the optimal number to predict saturated hydraulic conductivity at field scale.

Parasuraman et al. (2007) indicated that adoption of the ensemble technique in the formulation of PTFs helps in addressing one of the pertinent issues in any machine learning algorithm, namely generalization of the estimation results. In this study, 100 replicates were used to generate an ensemble of k-NN estimations. Using this number of replicates can be considered a safe choice in order to negate the impact of any single replicate on the final estimation results and obtain a high level of generalization of our results.

3.2. Optimizing the *k* and *p* terms

A next important preliminary step in establishing the k-NN PTF is the optimization of the two design parameters *k* and *p*. A gradual change of both parameters simultaneously will enable us to find an optimal combination of the *k* and *p* terms for the given task.

Figure 4 shows interdependence of the *k* and *p* terms and N_r, the number of samples in the reference dataset. Estimations developed from smaller data subsets (e.g. here $N_r = 100$ or 200) are more sensitive to changes in *k* and *p*. Including more samples from the reference dataset in each individual estimation (i.e. increasing *k*) beyond a threshold will generally yield worse estimations. This is because with small N_r, an increasing *k* will mean that a relatively large proportion of the dataset is included in the estimation, rather than a small, but more specific set of samples with very similar characteristics to the target sample. Hence, the estimates will tend to come closer and closer to the reference dataset mean, yielding less accurate 'local' estimates. This effect can be further enhanced by the choice of the *p* (weight) term, as best seen in Fig. 4a. The closer *p* is to zero, the more equal the weights are distributed among the chosen *k* number of samples. When *k* is relatively large, and *p* is kept small, even less similar samples will have a relatively large weight in the formulation of the final water retention estimate. On the contrary, the effect

of a relatively large p value is that even if more samples are used in the individual estimation (i.e. k is increased), the nearest samples (in their properties) would receive a very high proportion of the weights, while formulating the final estimate. In essence, a large p value can counteract the potentially negative effect of choosing a k value that is too large. This effect is best seen when k can be disproportionally high compared to N_r, as e.g. in Fig. 4a.

The above combined effect is less and less expressed with the increase of the size of the reference data set (N_r), at least within the examined range of k and p values. It is likely that following the above logic, with the further increase of k, we would see more impact of the choice of p on the estimation quality when larger N_r's are examined. Nevertheless, p should not be set too high either, since it carries the risk of giving too much weight to one or two individual samples, which may not best represent the characteristics of all similar samples. The simultaneous optimization of the k and p terms requires attentive consideration and good understanding of the underlying effects and consequences.

Based on Fig. 4, we tried to determine the k number which corresponds to the lowest RMSD (averaged through the eight matric potentials) for p values equal to 0.5, 1.0, 1.5, 2.0 and 2.5 for dataset sizes N_r equal to 100, 200, 300, 400, and 534 respectively. An average of all the optimal k numbers determined for p values equal to 0.5, 1.0, 1.5, 2.0, and 2.5 was calculated for each reference dataset size (Table 1). Since k can only be an integer, the calculated and rounded average k values found in Table 1 are plotted in Fig. 5 against the dataset size. An increasing trend with increasing dataset size was found and the best fitting equation relating the k number to the reference dataset size N_r was derived based on a power function:

551

552 $k = 0.724\ N_r^{0.468}$ [11]

553

554 Nemes et al. (2006a) found also a power function for the U.S. NRCS-SCS dataset (see
555 Eq. [6]). The derived equation yielded values of k very similar to the ones found by
556 Nemes et al. (2006a) for their dataset. Table 2 compares the k values derived from the
557 equation of Nemes et al. (2006a) and the ones derived from the equation found in this
558 paper. As noted above, values of k in their study and the present study are rounded to the
559 nearest integer, so the actual difference between k values may be even smaller.

560

561 To find the best combination between the k and p values, we compared the RMSDs
562 provided by each combination of k and p values for each reference dataset size using
563 contour plots (not shown here). The best p value was derived from the intersection
564 between the average k value given in Table 1 and the lowest RMSD (3 decimals
565 considered). We did not find a common trend for p value with the reference dataset size.
566 However for $N_r = 100$, $N_r = 400$ and $N_r = 534$, we found values around 1. Nemes et al.
567 (2006a) found that the p value ranged from 0.95 to 1.10. For $N_r = 200$ and $N_r = 300$, the
568 best p values were surprisingly close to 3.0 and 2.2 respectively which are quite large
569 values. However, even if a value of p around 1 were chosen for $N_r = 200$ and $N_r = 300$,
570 the RMSD increased by only 0.001 $m^3\ m^{-3}$, therefore, $p = 1$ seems to be a safe choice.
571 This is in line with the findings and recommendations by Nemes et al. (2006a) regarding
572 the relative insensitivity of the method to a range of p values. Because the difference and
573 its influence appears to be negligible, we decided to use the function previously used by
574 Nemes et al. (2006a) which relates the p value to the reference dataset size (see Eq. [7]).
575 Hence, a p value of 1.04 will correspond to the full dataset of 534 soil samples of the

IGBP-Trop dataset. This value is close to 1, which, following Nemes et al. (2006a), represents a simple inverse relationship between the weight and the distance of the selected sample. The generic settings of the *k* and *p* terms that were worked out for temperate-climate soils from the USA match closely with the optimal settings found for the IGBP-Trop dataset. In their study, Patil et al. (2012) also used the functions for *k* and *p* provided by Nemes et al. (2006a) and the reference dataset provided with the k-Nearest software (Nemes et al., 2008) and obtained good results for swelling-shrinking soils (RMSD < 0.05 m^3 m^{-3}).

3.3. Prediction of water retention from an international "tropical" database

In the present study, 14 combinations of input soil attributes were used to predict the eight water retention outputs. Table 3 gives a summary of the results in terms of MD, RMSD and R^2 at all the eight matric potentials, with the optimized settings and the various combinations of input parameters. The prediction performance of this k-NN algorithm is satisfactory in most cases. When considering individual MD, RMSD and R^2 values, we found: -0.009 m^3 m^{-3} < MD < 0.055 m^3 m^{-3}, 0.032 m^3 m^{-3} < RMSD < 0.087 m^3 m^{-3} and 0.280 < R^2 < 0.921. The average MD, RMSD and R^2 of eight matric potentials for each input variables combination was: 0.0066 m^3 m^{-3} < AvgMD < 0.0305 m^3 m^{-3}, 0.0439 m^3 m^{-3} < AvgRMSD < 0.0619 m^3 m^{-3} and 0.7010 < AvgR^2 < 0.8029. The RMSD values were situated between 0.051 and 0.063 m^3 m^{-3} for prediction of θ_{-10kPa} and between 0.032 and 0.038 m^3 m^{-3} for prediction of $\theta_{-1500kPa}$. These are encouraging results for these two points of the SWRC which are generally considered as good approximations of FC and PWP, respectively for soils in the humid tropics.

When focusing on the most basic predictor variables texture (SSC), BD and OC, generally used in hydraulic PTFs because of their availability in various soil survey

reports, it can be seen that the variation in RMSD values is particularly different when BD is included or not as a predictor (Table 3). A marked decreasing trend of RMSD values (from 0.076 m^3 m^{-3} to 0.033 m^3 m^{-3}) from the wet to the dry range of the SWRC can be observed when BD was not considered. On the contrary, when BD was included as predictor, RMSD values were low in the wet range (< 0.050 m^3 m^{-3}) followed by a slight increase in the intermediate range between matric potentials of -3 kPa and -50 kPa and again a decrease in the dry range (< 0.040 m^3 m^{-3}). In the intermediate range, the RMSD yielded by different combinations of inputs variables varies slightly with values between 0.050 m^3 m^{-3} and 0.060 m^3 m^{-3}. However, the contribution of BD as predictor to the slight decrease of the overall error in prediction at the intermediate range can still be observed. Vereecken et al. (2010) made similar observations regarding the evolution of RMSD values when a combination of SSC, BD and OM was used as predictors in the published PTFs considered in their review paper. The derived matric potentials by curve fitting (0, -50 and -250 kPa) did not show any out-of-pattern quality in the estimation of water retention. The results found in this study indicate that the performance of the k-NN algorithm is dependent on the matric potential at which water retention is predicted. Recently, Haghverdi et al. (2012) developed pseudo-continuous ANN PTFs for water retention. Notwithstanding the effect of different combinations of the aforementioned input variables, they also observed relatively large variations in RMSD values as a function of matric potential. For example, the RMSD values were 0.050 m^3 m^{-3} at -33 kPa and 0.035 m^3 m^{-3} at -1500 kPa. From Table 6 and from previous observations made by several authors such as Schaap et al. (2001), Vereecken et al. (2010) and Haghverdi et al. (2012), there seems to be an effect of the combinations of different input variables on the quality of prediction of water contents at various matric potentials. In the present study, the difference in prediction performance amongst models with the 14 input

627 variable combinations is more pronounced in the very wet range of the SWRC (at 0
628 and -1 kPa) with RMSD values between 0.038 and 0.087 m^3 m^{-3} and almost negligible at
629 the very dry range of the SWRC (at -250 and -1500 kPa) with RMSD values between
630 0.034 and 0.040 m^3 m^{-3}. In the intermediate range of the SWRC (from -3 to -50 kPa), the
631 RMSD values yielded by the 14 input combinations were approximately between 0.049
632 and 0.067 m^3 m^{-3} (Table 3). This can be explained by the major role played by soil
633 structure in the wet and in the intermediate ranges of the SWRC. Given that the best
634 proxy for soil structure in this study is BD, there will be a notable difference in prediction
635 performance between combinations including BD and combinations excluding BD as
636 input variable.
637
638 Table 3 further shows that the predictive ability of the k-NN algorithm in terms of bias
639 (MD), overall error (RMSD) and goodness-of-fit (R^2) closely depends on the
640 combination of the "predictors", i.e. the input attributes. Estimation quality may differ
641 significantly when one set of input attributes is used instead of another set. For example,
642 use of OC and pH were found to considerably reduce the quality of the prediction of
643 water retention in the wet range of the SWRC. When OC and pH are present in the input
644 attributes combination, they seem to favor soils in the training dataset which are quite
645 different from the target soil in their hydraulic behavior at the wet range of the SWRC.
646 On the other hand, they appeared to have a positive effect on the quality of the prediction
647 in the dry range of the SWRC. Likewise, BD contributes largely to the improvement of
648 the prediction of water retention in the wet range of the SWRC, while it is not the case in
649 the dry range. Besides soil texture which plays a major role in the whole range of the
650 SWRC, BD contributes largely to explaining water retention in the wet range of the
651 SWRC whereas OC is more influential in the dry range. The k-NN approach is thus able

652 to reflect this physical phenomenon. It was found that using the complete set of input
653 attributes i.e. SSC+BD+OC+pH+CEC was not the best option. As shown in Table 3, the
654 best combination appeared to be SSC+BD+CEC with the smallest bias error (AvgMD =
655 0.0066 m^3 m^{-3}), the smallest overall error (AvgRMSD = 0.0439 m^3 m^{-3}) and one of the
656 largest goodness-of-fit values (AvgR^2 = 0.8018), closely followed by the combination
657 SSC+BD (AvgMD = 0.0094 m^3 m^{-3}, AvgRMSD = 0.0444 m^3 m^{-3}, AvgR^2 = 0.8029). On
658 the other hand, the worst combination was found to be SSC+pH with the largest bias
659 error (AvgMD = 0.0305 m^3 m^{-3}), the largest overall error (AvgRMSD = 0.0619 m^3 m^{-3})
660 and the smallest goodness-of-fit value (AvgR^2 = 0.7010). One of the reasons of this result
661 could be the lack of a meaningful relationship between pH and water retention at all the
662 matric potentials in the test dataset with Pearson correlation coefficients $r < 0.203$.
663 Another reason could be the difference in distribution of pH values in the reference and
664 the test datasets (Fig. 2). In the reference dataset, the distribution of pH values is
665 somewhat skewed whereas in the test dataset, the pH values are normally distributed.
666 This suggests that pH will not be able to provide information necessary to identify the
667 most similar instances to a given target soil in relation with water retention. The variable
668 pH has thus a limited relationship with water retention and could worsen the prediction of
669 water retention particularly at high matric potentials, i.e. in the wet range of the SWRC,
670 at least using these particular datasets. Hodnett and Tomasella (2002) found that pH
671 contributed to the estimation of all four parameters of the van Genuchten (1980) equation
672 as it may be a crude indicator of the degree of weathering of soils in the tropics.
673
674 In their study on Vertisols, Patil et al. (2012) found that the inclusion of BD as predictor
675 in the k-NN technique led to a slight increase of the RMSD. They indicated that the BD
676 of Vertisols is known to change with soil water content (swelling-shrinking soils). This

677 particular behavior was observed and studied by various authors (e.g. Braudeau et al.,
678 2004; Cornelis et al., 2006).
679
680 Bulk density and CEC are good indirect indicators of the structure of the soil. Bulk
681 density gives an indication of total soil porosity, whereas CEC gives indications about the
682 clay mineralogy of the soil which is also responsible for the structural development and
683 porous behavior of the soil, besides retention of water by adsorption. Pachepsky and
684 Rawls (2003) indicated that BD is a measurable continuous variable which is indirectly
685 related to soil structure. In the same vein, Tranter et al. (2007) proposed a conceptual
686 model which considers BD as the result of particle packing and soil structure. Bronick
687 and Lal (2005) wrote that clay minerals influence properties that affect aggregation:
688 surface area, CEC, charge density, dispersivity and expandability. Based on CEC values,
689 a distinction can be made between soils with high activity clays (HAC) and soils with low
690 activity clays (LAC). Low activity clays such as kaolinite and halloysite generally occur
691 in highly weathered soils (e.g. Acrisols and Ferralsols), whereas HAC such as
692 montmorillonite are present in swelling-shrinking soils (e.g. Vertisols). As it is well
693 known, structure has a non-negligible influence on water retention at high matric
694 potentials. High CEC values are indications of soils with high water retention capacity
695 and poor internal drainage, whereas the opposite is true for soils with low CEC values.
696 Hodnett and Tomasella (2002) found that CEC can be a predictor of the van Genuchten
697 (1980) parameters as it may indicate the effect of mineralogy on water retention capacity
698 of soils in the tropics.
699
700 In the present study, the addition of OC seems not to improve significantly the prediction
701 compared to accounting for texture only. Similarly, Puckett et al. (1985) did not use

OM/OC as a predictor to derive water retention PTFs due to its low content in the soil samples from the Lower Coastal Plain in the USA. In their study on physical properties and moisture retention characteristics of some tropical soils in Nigeria, Lal (1978) did not find any effect of OM/OC on water retention. Zacharias and Wessolek (2007) suggested the exclusion of OM/OC as predictor in classic PTFs and proposed a new PTF that uses only physical properties such as soil texture and BD. On the contrary, Vereecken et al. (2010) observed that including OM/OC as predictor in "temperate" PTFs of e.g. Vereecken et al. (1989), Nemes et al. (2003) and Weynants et al. (2009) led to improved predictions, with the lowest RMSD values in the wet range and in the very dry range of the SWRC. This can be explained by the variability of OM/OC present in temperate and in tropical soils, with soils from temperate areas often having a substantial amount, and wider range of OM. This means that OM/OC can be a suitable predictor of water retention of soils in temperate regions. In contrast, OM/OC content is very low in the humid tropics due to a high rate of decomposition under high temperatures and abundant rainfall. Therefore, OM/OC may not have the variability to be an important variable in estimating the water retention for soils in the humid tropics.

Furthermore in Table 3, it is shown that the bias error (MD) can contribute, to various extents, to the overall error (RMSD). There is a clear trend to overestimate water retention in the wet and the middle range of the SWRC whereas there is a small but almost negligible trend to underestimate water retention at the dry range. The training dataset contains 80% of low activity clay (LAC) soils (i.e. with CEC < 20 cmol (+) kg^{-1} soil) and 20% of mixed activity clay (MIX) soils (i.e. with CEC between 20 and 62 cmol (+) kg^{-1} soil) whereas the test dataset contains more than 95% of LAC soils. While LAC soils are dominated by kaolinite and sesquioxides, MIX soils contain other clay minerals

727 such as montmorillonite which present a relatively higher water retention capacity than
728 kaolinite. Williams et al. (1983) observed that the presence of montmorillonite even in
729 quite small amounts in the soil samples was shown to be a discriminating property in
730 relation with water retention. In their evaluation study based on a limited test dataset of
731 soils from Lower Congo, Botula et al. (2012) found that the "temperate" PTFs of Gupta
732 and Larson (1979) largely overestimated the water retention of soils in the Lower Congo.
733 Botula et al. (2012) attributed this result to the differences in soil properties and in the
734 mineralogy between the test dataset and the dataset used to develop the PTFs. One
735 possible explanation of the large positive bias could be the difference in the distribution
736 of texture classes with a strong presence of silty soils in temperate (development) soil
737 datasets whereas clayey soils dominate in tropical (test) soil datasets. Another reason may
738 be the presence of montmorillonitic soils in the development dataset used by Gupta and
739 Larson (1979) and the large dominance of kaolinitic soils in the independent test dataset
740 used by Botula et al. (2012). In their study of the performance of various PTFs when
741 applied for Ferralsols from Cuba, Medina et al. (2002) indicated that clay type plays a
742 vital role in the retention and transmission properties of a given soil. It is the reason why
743 soils in the humid tropics can have much more clay than soils in the temperate regions
744 but a much lower water retention capacity.
745

746 **3.4. Prediction performance of the k-NN approach and the MLR approach**

747 The MLR PTFs of Hodnett and Tomasella (2002) use texture, BD, pH, OC and CEC as
748 predictors to estimate the van Genuchten parameters, whereas the point PTFs of Minasny
749 and Hartemink (2011) use texture, BD and OC as inputs. The RMSDs of these PTFs were
750 compared with the k-NN algorithm using different combinations of predictors: SSC+OC,

751 SSC+BD, SSC+BD+CEC as well as the full set of available predictors
752 (SSC+BD+OC+pH+CEC) (Table 4).
753
754 An independent one-sample t-test was run, evaluated at the 0.05 significance level, which
755 indicated that the RMSD values generated by the MLR PTFs and the k-NN models were
756 statistically different at each matric potential. The RMSDs of k-NN models varied by
757 matric potential and which set of predictors were used, but the PTFs of Hodnett and
758 Tomasella (2002) yielded comparable RMSD values to those of the k-NN algorithm with
759 certain combinations of inputs, primarily the SSC+BD and SSC+BD+CEC models. The
760 differences were rather small in most cases, but they were significant in all cases, given
761 the very small standard deviation of ensemble RMSDs. At near-saturation, the k-NN
762 estimates were more accurate, but in the intermediate matric potential range (from -10 to
763 -50 kPa) the Hodnett and Tomasella (2002) PTFs yielded smaller RMSD values than the
764 k-NN algorithm. The Hodnett and Tomasella (2002) PTFs and k-NN showed particularly
765 comparable performance in the dry range. We note that one of the points in the
766 intermediate range (i.e. -50 kPa) was derived by curve fitting for the Lower Congo data
767 set, which may have introduced some degree of extra uncertainty into the estimations.
768 The point PTFs of Minasny and Hartemink (2011) gave significantly greater RMSD
769 values than the PTFs of Hodnett and Tomasella (2002) and any of the examined k-NN
770 algorithms at the two available matric potentials (Table 4).
771
772 Any direct comparison of the performance of PTFs that do not use the same inputs is
773 influenced by the cost and benefit of any extra variable(s), so conclusions have to be
774 drawn carefully. The k-NN algorithm that uses SSC+BD+OC+pH+CEC requires the
775 same input attributes as the PTF of Hodnett and Tomasella (2002) that predicts the van

776 Genuchten (1980) parameters. On the other hand, the k-NN algorithm using SSC+BD
777 uses the same inputs as the -10 kPa PTF of Minasny and Hartemink (2011), while the k-
778 NN algorithm using SSC+OC uses the same inputs as the -1500 kPa PTF of Minasny and
779 Hartemink (2011). In our comparison with the two MLR models, it can be concluded that
780 the presented k-NN models that use the same inputs, show better performance measures
781 than the Minasny and Hartemink (2011) PTFs. On the other hand, when e.g. the SSC+BD
782 k-NN model is compared to the Hodnett and Tomasella (2002) PTFs, a somewhat weaker
783 performance is achieved, but with significantly smaller number of inputs – i.e. k-NN did
784 not use OC, pH and CEC as inputs. It is of particular value in data- and resource-poor
785 environments if the need for input is minimized in a quest to obtain estimates of
786 expensive but important soil hydraulic properties. The Hodnett and Tomasella (2002)
787 PTFs require the user to have all five of the above listed properties available in order to
788 estimate water retention of a tropical soil, which can be a serious limitation in their
789 applicability. The presented k-NN approach can be used in a hierarchical way, adjusting
790 the used inputs to their availability, and acceptably good and stable estimation results can
791 already be achieved by using only texture and bulk density as predictors. Among the
792 examined PTFs, the presented k-NN based PTFs introduced in this paper appear to show
793 the best value, when statistical performance is combined with the PTFs' need for input.
794 Given that the source of the development data was the same for the two MLR and the k-
795 NN PTFs, it is likely that the PTF development methodology and the data they have been
796 tested on are the combined reason for that finding. Given its capability and flexibility in
797 utilizing limited or a wider range of predictors hierarchically, based on their availability,
798 the k-NN technique presents far greater number of choices and flexibility to the user than
799 published MLR PTFs do. Additionally, given that all calculations are made real-time in
800 k-NN, as growth and development of tropical soil databases is expected, those new data

801 can be taken into account by the k-NN technique without the need to redevelop any
802 equations, which would be necessary with MLR PTFs like the ones of Hodnett and
803 Tomasella (2002) and Minasny and Hartemink (2011).
804
805 In preparation for future needs and increased computing capabilities, the k-NN technique
806 can also readily provide an estimate of the uncertainty when ensembles of estimations are
807 generated. Such advances can be well taken into account while parameterizing
808 simulation-based environmental risk-assessment and scenario studies. The presented k-
809 NN application also demonstrated how any number of points can be estimated
810 simultaneously on the SWRC curve, given that those points exist in the source database.
811 Therefore, besides its capability to provide SWRC estimates of competitive quality, the
812 proposed k-NN approach gives a number of additional benefits to the user, compared to
813 existing MLR approaches. When provided with an enhanced user interface, similar in
814 nature to the k-Nearest software of Nemes et al. (2008), the k-NN variant developed in
815 this paper can be easily implemented by potential users interested in soils of the humid
816 tropics.

4. Conclusions

A variant of the k-NN algorithm developed by Nemes et al. (2006a) has been applied and tested to predict water retention of soils from the Lower Congo in Central Africa based on an international dataset (IGBP-Trop) of soils of the (sub)humid tropics. Two design-parameters *k* and *p* that are user-defined and determined before and independent of applying the non-parametric k-NN algorithm were optimized to better take advantage of the k-NN variant introduced in this study. The optimized *k* and *p* values were found to be similar to those of previous studies. The results showed that this k-NN variant was able to estimate water retention at eight different matric potentials (0, -1, -3, -10, -20, -50, -250 and -1500 kPa), i.e. from the wet to the dry range of the SWRC with an average RMSD < 0.046 m^3 m^{-3} when SSC+BD or SSC+BD+CEC were selected as input variables. The overall prediction performance of the proposed non-parametric approach was compared with two tropical equation-based PTFs of Hodnett and Tomasella (2002) and Minasny and Hartemink (2011) based on the MLR approach. The results suggest that the k-NN approach shows comparable prediction performance to the examined MLR PTFs, which makes it a competitive alternative to those equations-based PTFs that are currently available to predict water retention of soils in the humid tropics. While performing similarly, the presented k-NN variant provides a great degree of flexibility and extra options to the user. The user can, for example, (1) incorporate additional data by appending to or replacing the reference database without the need or burden of redeveloping new equations, (2) develop the estimations real-time, decide real-time what inputs to use and vary them from sample to sample if desired, (3) estimate any number and combination of SWRC points simultaneously, driven by their availability in the reference/development dataset, and (4) generate an uncertainty measure to the estimates.

These advantages can be particularly beneficial in the context of developing countries where there is growing demand – as well as potential – to continuously develop soil databases - and subsequent simulation-based studies - for pedological, agricultural and environmental studies. For future research, we recommend testing the ability of this technique to predict water retention of other soils found in the tropics, for example volcanic soils that present some specific properties. These soils present a completely different mineralogy than highly weathered soils or swelling-shrinking soils and may need a completely different reference/training dataset than the IGBP-Trop dataset to provide acceptable estimations of their hydraulic characteristics.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that improved the quality of the paper.

5. References

Babalola, O. 1979. Spatial variability of soil water properties for a tropical soil of Nigeria. Soil Sci. 126:269-279.

Bannayan, M., and G. Hoogenboom. 2009. Using pattern recognition for estimating cultivar coefficients of a crop simulation model. Field Crop. Res. 111:290-302.

Botula, Y.-D., W.M. Cornelis, G. Baert, and E. Van Ranst. 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (D.R. Congo). Agric. Water Manage. 111:1-10.

Bouma, J. 1989. Using soil survey data for quantitative land evaluation. Adv.Soil Sci. 9:177-213.

Bouma, J., and J.A.J. van Lanen. 1987. Transfer functions and threshold values: From soil characteristics to land qualities. In/ Quantified land evaluation. Proc. Worksh. ISSS and SSSA, Washington, DC. 27 Apr.–2 May 1986. K.J. Beek et al. (eds). Int. Inst. Aerospace Surv. Earth Sci. Publ. No. 6. ITC Publ. Enschede, the Netherlands, pp. 106-110.

Braudeau, E., J.P. Frangi, and R.H. Mohtar. 2004. Characterizing nonrigid aggregated soil-water medium using its shrinkage curve. Soil Sci. Soc. Am. J. 68:359-370.

Bronick, C.J., and R. Lal. 2005. Soil structure and management: a review. Geoderma 124:3-22.

Brooks, R.H., and A.T. Corey. 1964. Hydraulic properties of porous media. Hydrology Paper 3, Colorado State University, Fort Collins, Colorado, USA.

Buishand, T.A., and T. Brandsma. 2001. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. Water Resour. Res. 37:2761-2776.

Cornelis, W.M., M. Khlosi, R. Hartmann, M. Van Meirvenne, and B. De Vos. 2005. Comparison of unimodal analytical expressions for the soil-water retention curve. Soil Sci. Soc. Am. J. 69:1902-1911

Cornelis, W.M., J. Corluy, H. Medina, J. Diaz, R. Hartmann, M. Van Meirvenne, and M.E. Ruiz. 2006. Measuring and modelling the soil shrinkage characteristic curve. Geoderma 137:179-191.

Dasarathy, B.V. (ed.) 1991. Nearest neighbor (NN) Norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos, CA.

Deng, H.L., 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. doi:10.1029/2008wr007477.

Elshorbagy, A.A., G. Corzo, S. Srinivasulu, and D. Solomatine. 2010a. Experimental investigation of the predictive capabilities of soft computing techniques in hydrology- Part I: Concepts and methodology. Hydrol. Earth Syst. Sc. 14:1931-1941.

Elshorbagy, A.A., G. Corzo, S. Srinivasulu, and D. Solomatine. 2010b. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part II: Application. Hydrol. Earth Syst. Sc. 14:1943-1961.

Finke, P.A., J.H.M. Wösten, and M.J.W. Jansen. 1996. Effects of uncertainty in major input variables on simulated functional soil behaviour. Hydrol. Process. 10:661-669.

Gharahi Ghehi, N., A. Nemes, A. Verdoodt, W.M. Cornelis, E. Van Ranst, and P. Boeckx. 2012. Use of the Nonparametric nearest neighbor and boosted regression tree techniques to estimate soil bulk density in tropical rainforest soils. Soil Sci. Soc. Am. J. 76:1172-1183.

Givi, J., S.O. Prasher, and R.M. Patel. 2004. Evaluation of pedotransfer functions in predicting the soil water contents at field capacity and wilting point. Agric. Water Manage. 70:83-96.

Guber, A.K., Y.A. Pachepsky, M.Th. van Genuchten, W.J. Rawls, J. Simunek, D. Jacques, T.J. Nicholson, and R.E. Cady. 2006. Field-scale water flow simulations using ensembles of pedotransfer functions for soil water retention. Vadose Zone J. 5:234-247.

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.

Haghverdi, A., W.M. Cornelis, and B. Ghahraman. 2012. A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. J Hydrol. http://dx.doi.org/10.1016/j.jhydrol.2012.03.036 (in press).

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: prediction, inference and data mining. 2nd ed. Springer Verlag, New York.

Haykin, S. 1994. Neural Networks, a comprehensive foundation. 1st ed. Macmillan College Publishing Company, New York.

Hecht-Nielsen, R. 1990. Neurocomputing. Addison-Wesley, Reading, MA.

Hodnett, M.G., and J. Tomasella. 2002. Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: a new water-retention pedo-transfer functions developed for tropical soils. Geoderma 108:155-180.

Hopmans, J.W., J. Simunek, N. Romano, and W. Durner. 2002. Water retention and storage: Inverse methods. In: J.H. Dane & G.C. Topp (eds), Methods of soil analysis: Part 4-Physical methods. SSSA Book Series N° 5. SSSA, Madison, WI: 963-1004.

IUSS Working Group WRB. 2006. World Reference Base for Soil Resources 2006, 2nd ed. World Soil Resources Reports No. 103. FAO, Rome.

Jagtap, S.S., U. Lall, J.W. Jones, A.J. Gijsman, and J.T. Ritchie. 2004. Dynamic nearest-neighbor method for estimating soil water parameters. T. ASAE 47:1437-1444.

Köhn, M. 1929. Korngrößenanalyse vermittels Pipettanalyse. Tonindustrie-Zeitung 5: 729-731.

Lal, R. 1978. Physical-properties and moisture retention characteristics of some nigerian soils. Geoderma 21:209-223.

Loosvelt, L., V.R.N. Pauwels, W.M. Cornelis, G.J.M. De Lannoy, and N.E.C. Verhoest. 2011. Impact of soil hydraulic parameter uncertainty on soil moisture modeling. Water Resour. Res. 47. doi:10.1029/2010wr009204.

Medina, H., M. Tarawally, A. del Valle, and M.E. Ruiz. 2002. Estimating soil water retention curve in rhodic ferralsols from basic soil data. Geoderma 108:277-285.

Minasny, B., and A.E. Hartemink. 2011. Predicting soil properties in the tropics. Earth-Sci. Rev. 106:52-62.

Moeys, J., M. Larsbo, L. Bergstrom, C.D. Brown, Y. Coquet, and N.J. Jarvis. 2012. Functional test of pedotransfer functions to predict water flow and solute transport with the dual-permeability model MACRO. Hydrol. Earth Syst. Sci. 16:2069-2083. doi:10.5194/hess-16-2069-2012.

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

Mucherino, A., P. Papajorgji, and P.M. Pardalos. 2009. Data mining in agriculture. Springer, New York.

Nemes, A., J.H.M. Wosten, A. Lilly, and J.H. Oude Voshaar. 1999. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. Geoderma 90:187-202.

Nemes, A., M.G. Schaap, and J.H.M. Wosten. 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Sci. Soc. Am. J. 67:1093-1102.

Nemes, A., W.J. Rawls, and Y.A. Pachepsky. 2006a. Use of the non-parametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70:327-336.

Nemes, A., W.J. Rawls, Y.A. Pachepsky, and M.Th. van Genuchten. 2006b. Sensitivity of the nearest neighbor approach to estimate soil hydraulic properties. Vadose Zone J. 5:1222-1235.

Nemes, A., R.T. Roberts, W.J. Rawls, Y.A. Pachepsky, and M.Th. van Genuchten. 2008. Software to estimate-33 and-1500 kPa soil water retention using the non-parametric k-Nearest Neighbor technique. Environ. Modell. Soft. 23:254-255.

Nemes, A., Y.A. Pachepsky, and D.J Timlin. 2011. Toward improving global estimates of field soil water capacity. Soil Sci. Soc. Am. J. 75:807-812.

Noble, W.S. 2006. What is a support vector machine? Nat. Biotechnol. 24:1565-1567.

Ottoni Filho, T.B., and M.V. Ottoni. 2010. A variation of the Field Capacity (FC) definition and a FC database for Brazilian soils. 19th World Congress of Soil Science, Soil Solutions for a Changing World 1-6 August 2010, Brisbane, Australia. Published on DVD.

Pachepsky, Y.A., and W.J. Rawls. 2003. Soil structure and pedotransfer functions. Eur. J. Soil Sci. 54:443-451.

Pachepsky, Y.A., W.J. Rawls, and H.S. Lin. 2006. Hydropedology and pedotransfer functions. Geoderma 131:308-316.

Parasuraman, K., A. Elshorbagy, and B.C. Si. 2006. Estimating saturated hydraulic conductivity in spatially variable fields using neural network ensembles. Soil Sci. Soc. Am. J. 70:1851-1859.

Parasuraman, K., A. Elshorbagy, and B.C. Si. 2007. Estimating saturated hydraulic conductivity using genetic programming. Soil Sci. Soc. Am. J. 71:1676-1684.

Patil, N.G., D.K. Pal, C. Mandal, and D.K. Mandal. 2012. Soil water retention characteristics of vertisols and pedotransfer functions based on nearest neighbor and neural networks approaches to estimate AWC. J. Irrig. Drain. E-ASCE 138:177-184.

Perkins, K., and J. Nimmo. 2009. High-quality unsaturated zone hydraulic property data for hydrologic applications. Water Resour. Res. 45. W07417, doi:10.1029/2008WR007497.

Pidgeon, J.D. 1972. The measurement and prediction of available water capacity of ferralitic soils in Uganda. J. Soil Sci. 23:431-441.

Puckett, W.E., J.H. Dane, and B.F. Hajek. 1985. Physical and mineralogical data to determine soil hydraulic properties. Soil Sci. Soc. Am. J. 49:831-836.

Rajkai, K., S. Kabos, and M.Th. van Genuchten. 2004. Estimating the water retention curve from soil properties: comparison of linear, nonlinear and concomitant variable methods. Soil Till. Res. 79:145-152.

Rawls, W.J., and D.L. Brakensiek. 1982. Estimating soil water retention from soil properties. J. Irrig. Drainage Div. ASCE 108:166-171.

Reichardt, K. 1988. Capacidade de campo. Rev. Bras. Ci. Solo 12:211-216.

Reichert, J.M., J.A. Albuquerque, D.R. Kaiser, D.J. Reinert, F.L. Urach, and R. Carlesso. 2009. Estimation of water retention and availability in soils of Rio Grande do Sul. Rev. Bras. Ci. Solo 33:1547-1560.

Schaap, M.G. 2005. Models for indirect estimation of soil hydraulic properties. In: M. Anderson (ed), Encyclopedia of hydrological sciences. John Wiley & Sons, Ltd.

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.

Schwartz, R.C., and S.R. Evett. 2002. Estimating hydraulic properties of a fine-textured soil using a disc infiltrometer. Soil Sci. Soc. Am. J. 66:1409-1423.

Sharma, M.L., and G. Uehara. 1968. Influence of soil structure on water relations in low humic latosols . I. Water retention. Soil Sci. Soc. Am. P. 32:765-770.

Soil Survey Staff. 1975. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. USDA Handb. 436. U.S. Gov. Print. Office. Washington, DC.

Tempel, P., N. H. Batjes, and V. W. P. van Engelen. 1996. IGBP-DIS soil data set for pedotransfer function development, Working Paper and Preprint 96/05, Int. Soil Ref. and Inf. Cent. (ISRIC), Wageningen, Netherlands.

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.

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.

Tranter, G., B. Minasny, A.B. McBratney, B. Murphy, N.J. McKenzie, M. Grundy, and D. Brough. 2007. Building and testing conceptual and empirical models for predicting soil bulk density. Soil Use and Manage 23:437-443.

Twarakavi, N.K.C., M. Sakai, and J. Simunek. 2009. An objective analysis of the dynamic nature of field capacity. Water Resour. Res. 45. doi:10.1029/2009wr007944.

USDA. 1951. Soil survey manual. U.S. Dep. Agric. Handb. No. 18. U.S. Gov. Print Office, Washington, DC.

van Dam, J.C., J.N.M. Stricker, and P. Droogers. 1994. Inverse method to determine soil hydraulic functions from multi-step outflow experiments. Soil Sci. Am. J. 58:647-652.

van den Berg, M., E. Klamt, L.P. vanReeuwijk, and W.G. Sombroek. 1997. Pedotransfer functions for the estimation of moisture retention characteristics of Ferralsols and related soils. Geoderma 78:161-180.

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.

Van Ranst, E., M. Verloo, A. Demeyer, and J.M. Pauwels. 1999. Manual for the soil chemistry and fertility laboratory. Analytical methods for soils and plants. Equipment and management of consumables. International Training Centre for Post-Graduate Soil Scientists, Universiteit Gent, Gent, Belgium.

Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V. 1998. Statistical Learning Theory. John Wiley & Sons, New York.

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.

Vereecken, H. 1995. Estimating the unsaturated hydraulic conductivity from theoretical-models using simple soil properties. Geoderma 65:81-92.

Vereecken, H., M. Weynants, M. Javaux, Y. Pachepsky, M.G. Schaap, and M.Th. van Genuchten. 2010. Using pedotransfer functions to estimate the van genuchten-mualem soil hydraulic properties: A review. Vadose Zone J. 9:795-820.

Weynants, M., H. Vereecken, and M. Javaux. 2009. Revisiting Vereecken pedotransfer functions: Introducing a closed-form hydraulic model. Vadose Zone J. 8:86-95.

Williams, J., R.E. Prebble, W.T. Williams, and C.T. Hignett. 1983. The influence of texture, structure and clay mineralogy on the soil-moisture characteristic. Aust. J. Soil Res. 21:15-32.

Wösten, J.H.M., A. Lilly, A. Nemes, and C. Le Bas. 1999. Development and use of a database of hydraulic properties of European soils. Geoderma 90:169-185.

Zacharias, S., and G. Wessolek. 2007. Excluding organic matter content from pedotransfer predictors of soil water retention. Soil Sci. Soc. Am. J. 71:43-50.

LIST OF FIGURE CAPTIONS

Fig. 1. Variation of clay, silt and sand in the IGBP-Trop (circles) and the Lower Congo soil datasets (crosses).

Fig. 2. Box-plots of some physical and chemical properties of the soils of (1) IGBP-Trop (reference dataset) and (2) the Lower Congo (test dataset). BD is bulk density (Mg m^{-3}), OC is organic carbon content (%) and CEC is cation exchange capacity (cmol kg^{-1} soil).

Fig. 3. Running root mean squared differences (RMSDs) for the Lower Congo test dataset for up to 100 ensembles using sand, silt, clay, bulk density organic carbon, pH and cation exchange capacity as input attributes and water retention at (a) -1 kPa, (b) -20 kPa and (c) -1500 kPa as output attributes.

Fig. 4.Variations of the root mean squared differences (RMSDs) with the number of nearest neighbors k in function of p values and reference dataset sizes N_r.

Fig. 5. Effect of dataset size on the optimal choice of the number of selected neighbors.

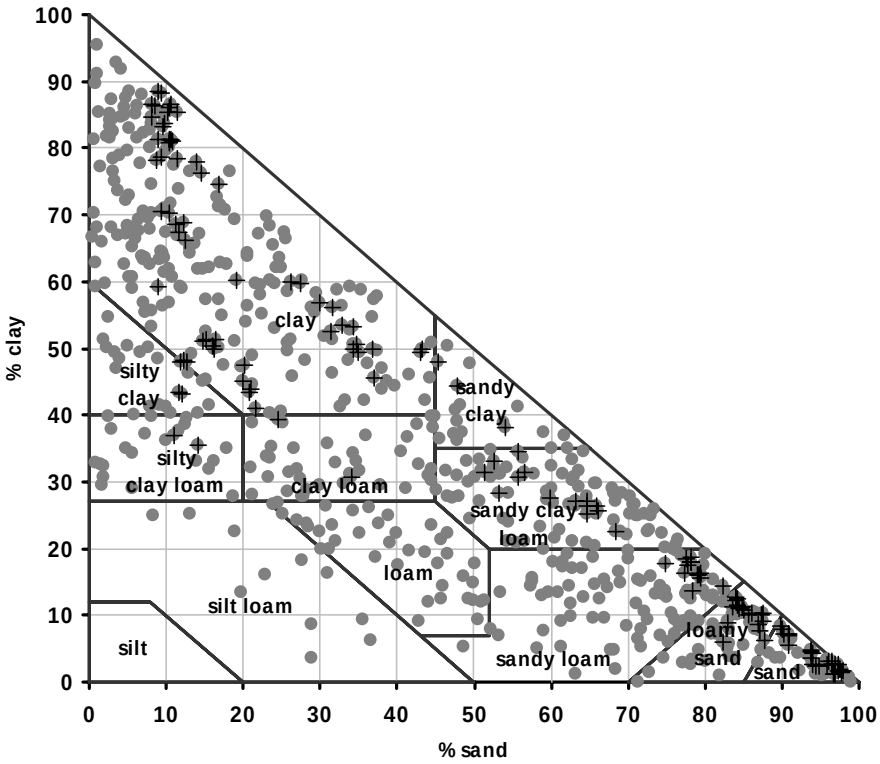

Fig. 1. Variation of clay, silt and sand in the IGBP-Trop (circles) and the Lower Congo soil datasets (crosses).

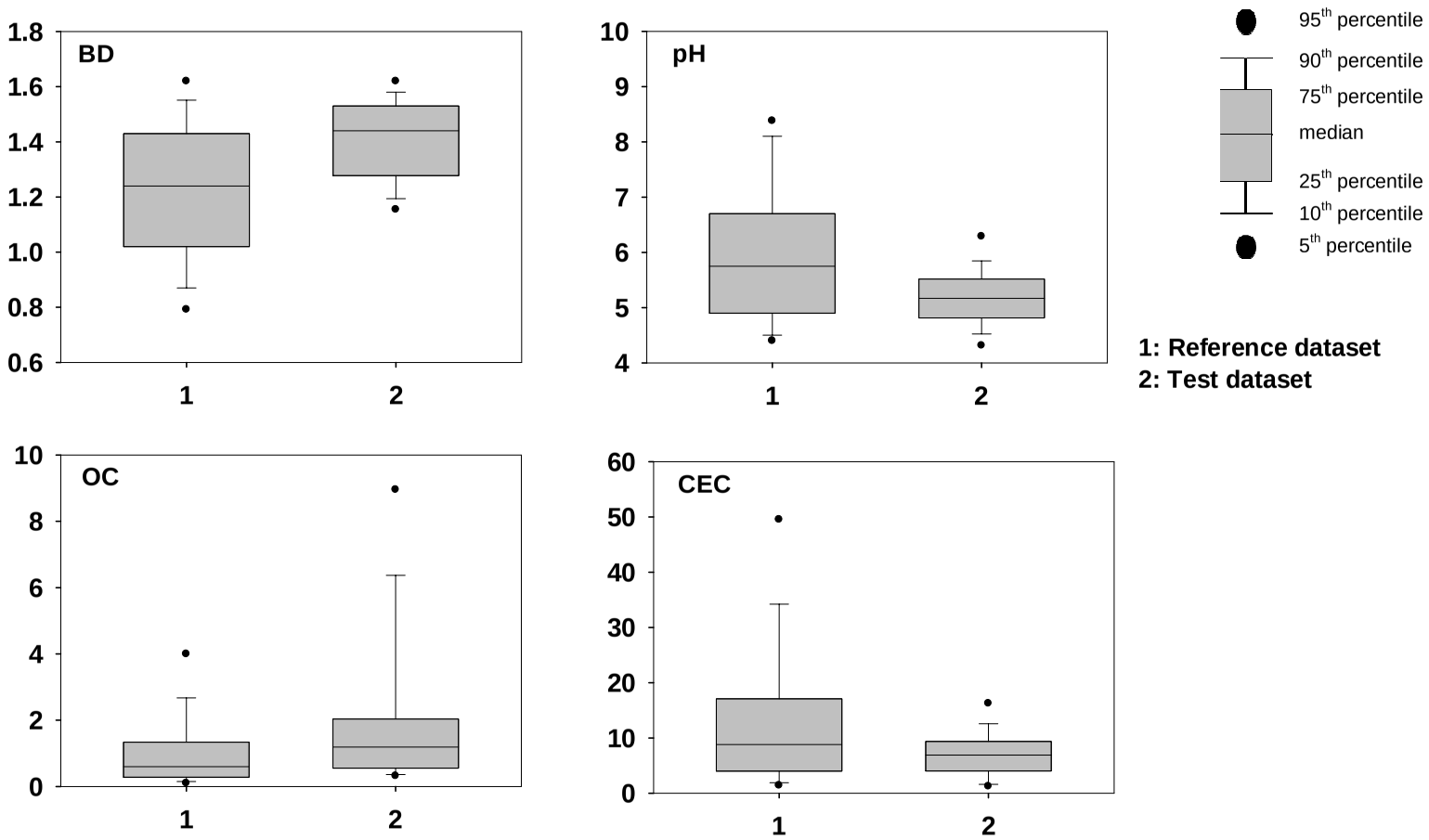

Fig. 2. Box-plots of some physical and chemical properties of the soils of (1) IGBP-Trop (reference dataset) and (2) the Lower Congo (test dataset). BD is bulk density (Mg m^{-3}), OC is organic carbon content (%) and CEC is cation exchange capacity (cmol kg^{-1} soil).

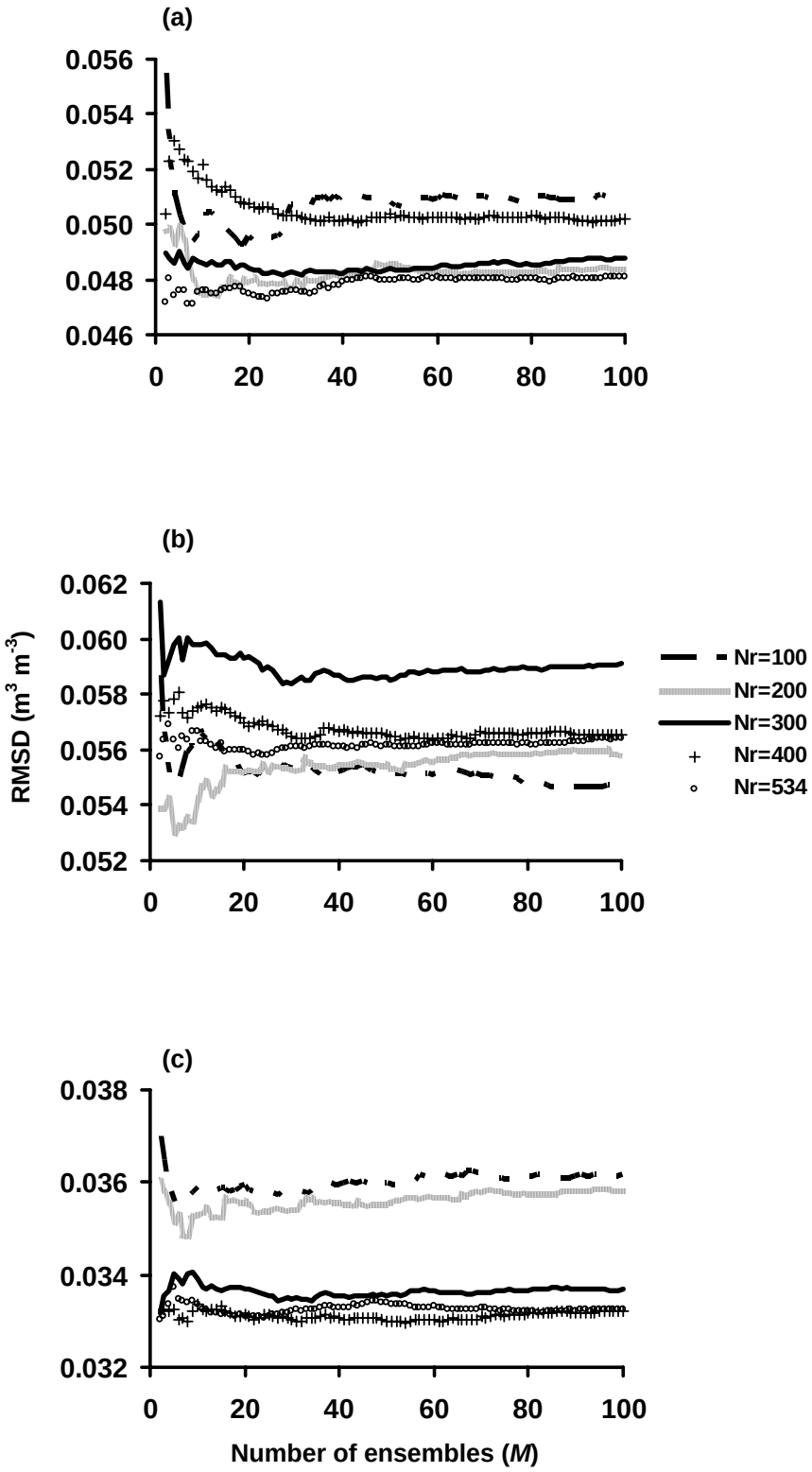

Fig. 3. Running root mean squared differences (RMSDs) for the Lower Congo test dataset for up to 100 ensembles using sand, silt, clay, bulk density, organic carbon, pH and cation exchange capacity as input attributes and water retention at (a) -1 kPa, (b) -20 kPa and (c) -1500 kPa as output attributes.

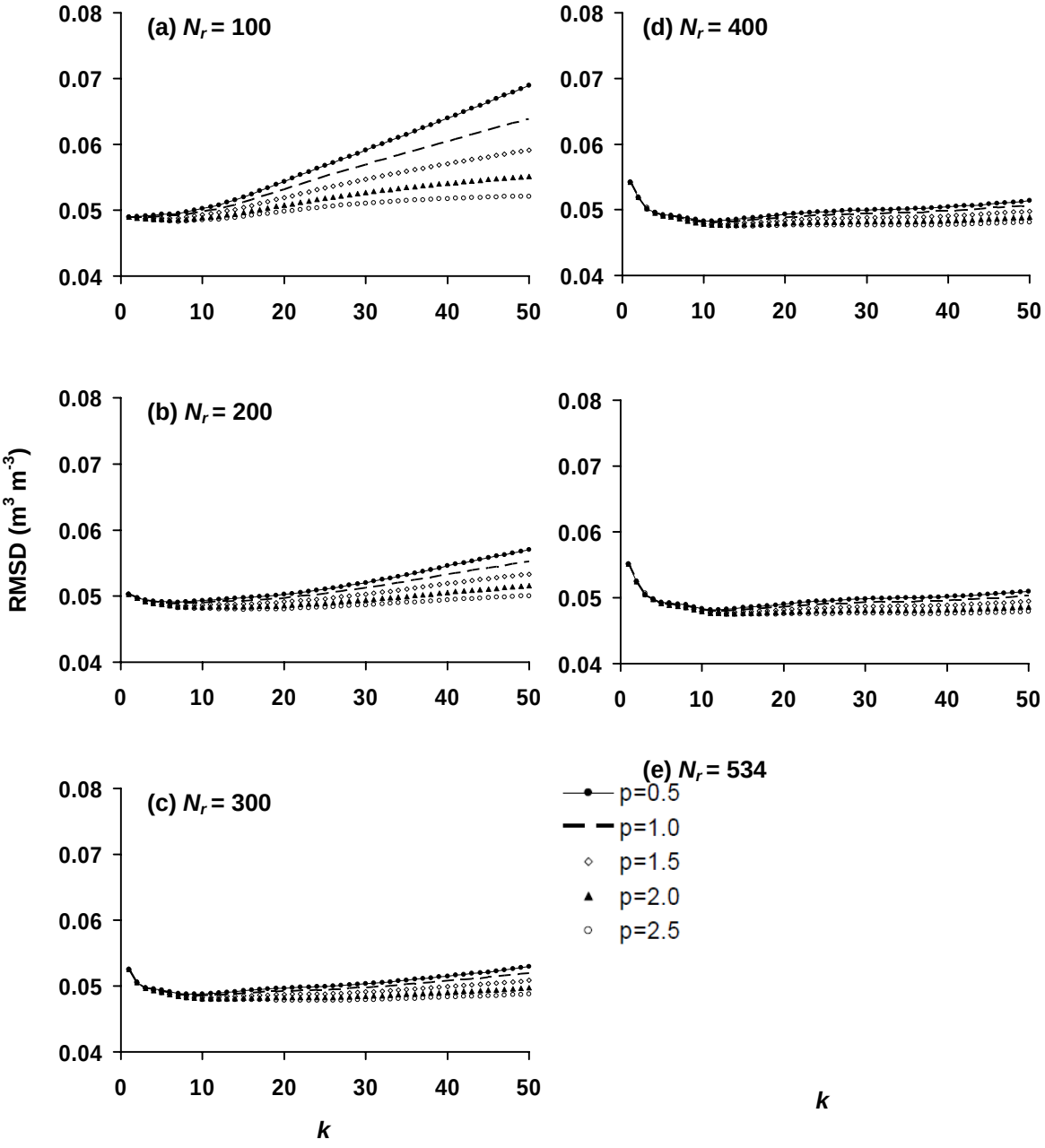

Fig. 4. Variations of the root mean squared differences (RMSDs) with the number of nearest neighbors k in function of p values and reference dataset sizes N_r.

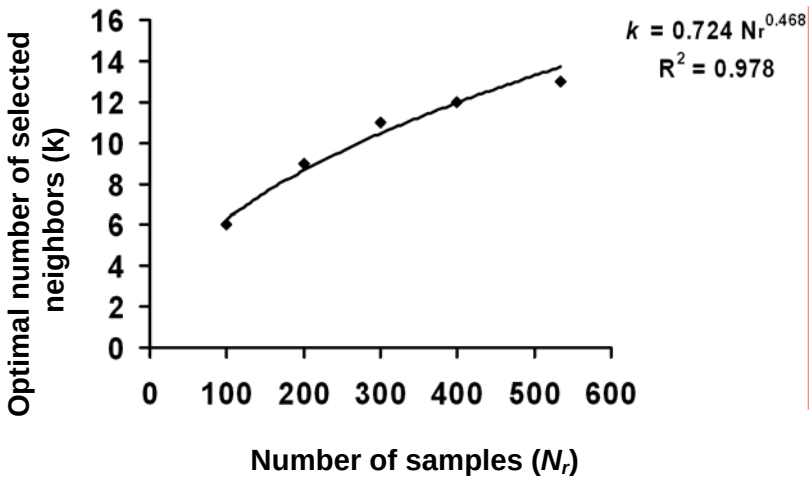

Fig. 5. Effect of dataset size on the optimal choice of the number of selected neighbors.

TABLES

Table 1. Number of nearest neighbors (k) corresponding to the lowest RMSD for different values of p and different dataset sizes N_r.

	N_r=100	N_r=200	N_r=300	N_r=400	N_r=534
p			***k***		
0.5	1†	7	9	11	11
1.0	2†	7	10	11	12
1.5	4	7	10	11	13
2.0	7	10	10	13	13
2.5	7	15	14	13	14
Average‡	**6**	**9**	**11**	**12**	**13**

† These values were not taken into account in the calculation of the average k because they did not correspond to a global or a local minimum for RMSD.
‡ Average values are rounded to the nearest integer.

Table 2. Comparison of the *k* number generated by the power function of Nemes et al. (2006a) and the power function derived for this study.

N_r	*k* calculated from Nemes et al. (2006a) function	*k* calculated from the present function
100	6	6
200	9	9
300	11	10
400	13	12
534	14	14

Table 3. Summary of results in terms of MD, RMSD and R^2, for the k-NN method with optimized settings at eight different matric potentials and using 14 combinations of input attributes.†

	Predicted water content								
Input attributes	θ_{0kPa}	θ_{-1kPa}	θ_{-3kPa}	θ_{-10kPa}	θ_{-20kPa}	θ_{-50kPa}	$\theta_{-250kPa}$	$\theta_{-1500kPa}$	
	MD (m^3 m^{-3})								**AvgMD**
SSC	0.039 (0.0025)	0.037 (0.0027)	0.013 (0.0033)	0.013 (0.0021)	0.029 (0.0023)	0.032 (0.0022)	-0.004 (0.0018)	-0.005 (0.0016)	**0.0193**
SSC+BD	0.013 (0.0018)	0.014 (0.0020)	-0.006 (0.0031)	0.003 (0.0022)	0.022 (0.0022)	0.027 (0.0021)	0.002 (0.0015)	0.000 (0.0014)	**0.0094**
SSC+OC	0.048 (0.0026)	0.043 (0.0026)	0.017 (0.0030)	0.016 (0.0027)	0.032 (0.0027)	0.035 (0.0025)	-0.005 (0.0018)	-0.006 (0.0017)	**0.0225**
SSC+BD+OC	0.021 (0.0021)	0.019 (0.0022)	-0.002 (0.0030)	0.006 (0.0024)	0.024 (0.0023)	0.028 (0.0022)	-0.001 (0.0017)	-0.003 (0.0016)	**0.0115**
SSC+pH	0.053 (0.0031)	0.052 (0.0031)	0.021 (0.0037)	0.029 (0.0033)	0.044 (0.0032)	0.046 (0.0030)	-0.001 (0.0017)	0.000 (0.0016)	**0.0305**
SSC+CEC	0.038 (0.0027)	0.036 (0.0028)	0.011 (0.0030)	0.011 (0.0025)	0.027 (0.0026)	0.030 (0.0024)	-0.005 (0.0017)	-0.004 (0.0017)	**0.0180**
SSC+pH+CEC	0.049 (0.0027)	0.048 (0.0026)	0.017 (0.0030)	0.026 (0.0029)	0.042 (0.0028)	0.044 (0.0026)	0.000 (0.0018)	0.001 (0.0017)	**0.0284**
SSC+OC+pH	0.055 (0.0023)	0.051 (0.0022)	0.019 (0.0027)	0.027 (0.0026)	0.042 (0.0026)	0.044 (0.0024)	-0.001 (0.0020)	0.000 (0.0019)	**0.0296**
SSC+OC+CEC	0.049 (0.0029)	0.044 (0.0028)	0.016 (0.0029)	0.014 (0.0028)	0.029 (0.0028)	0.032 (0.0026)	-0.008 (0.0018)	-0.008 (0.0017)	**0.0210**
SSC+BD+pH	0.017 (0.0018)	0.018 (0.0018)	-0.005 (0.0027)	0.012 (0.0026)	0.029 (0.0026)	0.034 (0.0024)	0.002 (0.0020)	0.004 (0.0019)	**0.0139**
SSC+BD+CEC	0.012 (0.0018)	0.013 (0.0020)	-0.009 (0.0030)	-0.001 (0.0021)	0.018 (0.0021)	0.023 (0.0020)	-0.002 (0.0017)	-0.001 (0.0016)	**0.0066**
SSC+BD+pH+CEC	0.016 (0.0018)	0.018 (0.0019)	-0.005 (0.0028)	0.011 (0.0026)	0.029 (0.0026)	0.033 (0.0025)	0.002 (0.0022)	0.003 (0.0020)	**0.0134**
SSC+OC+pH+CEC	0.053 (0.0025)	0.050 (0.0023)	0.018 (0.0026)	0.026 (0.0026)	0.041 (0.0026)	0.044 (0.0024)	-0.001 (0.0020)	0.000 (0.0019)	**0.0289**
SSC+BD+OC+pH+CEC	0.023 (0.0021)	0.023 (0.0020)	-0.001 (0.0026)	0.015 (0.0025)	0.032 (0.0025)	0.036 (0.0023)	0.000 (0.0021)	0.001 (0.0019)	**0.0161**
	RMSD (m^3 m^{-3})								**AvgRMSD**
SSC	0.076 (0.0017)	0.070 (0.0017)	0.060 (0.0022)	0.057 (0.0016)	0.055 (0.0019)	0.054 (0.0019)	0.037 (0.0009)	0.035 (0.0008)	**0.0555**
SSC+BD	0.039 (0.0014)	0.039 (0.0014)	0.050 (0.0020)	0.051 (0.0013)	0.050 (0.0016)	0.051 (0.0017)	0.039 (0.0010)	0.036 (0.0008)	**0.0444**
SSC+OC	0.078 (0.0022)	0.070 (0.0021)	0.057 (0.0021)	0.058 (0.0018)	0.057 (0.0024)	0.056 (0.0023)	0.034 (0.0012)	0.033 (0.0013)	**0.0554**
SSC+BD+OC	0.048 (0.0024)	0.045 (0.0021)	0.051 (0.0023)	0.052 (0.0016)	0.050 (0.0017)	0.050 (0.0017)	0.034 (0.0010)	0.033 (0.0010)	**0.0454**
SSC+pH	0.086 (0.0023)	0.078 (0.0021)	0.058 (0.0020)	0.063 (0.0026)	0.067 (0.0033)	0.067 (0.0032)	0.039 (0.0007)	0.037 (0.0007)	**0.0619**
SSC+CEC	0.075 (0.0023)	0.071 (0.0022)	0.060 (0.0024)	0.057 (0.0016)	0.056 (0.0020)	0.055 (0.0020)	0.040 (0.0010)	0.038 (0.0009)	**0.0565**
SSC+pH+CEC	0.083 (0.0022)	0.076 (0.0020)	0.058 (0.0021)	0.060 (0.0019)	0.063 (0.0025)	0.063 (0.0024)	0.039 (0.0008)	0.038 (0.0009)	**0.0600**
SSC+OC+pH	0.087 (0.0019)	0.077 (0.0017)	0.058 (0.0019)	0.062 (0.0019)	0.064 (0.0023)	0.063 (0.0022)	0.037 (0.0012)	0.035 (0.0013)	**0.0604**
SSC+OC+CEC	0.080 (0.0021)	0.072 (0.0019)	0.058 (0.0020)	0.058 (0.0017)	0.056 (0.0024)	0.054 (0.0023)	0.035 (0.0010)	0.033 (0.0011)	**0.0558**
SSC+BD+pH	0.045 (0.0016)	0.042 (0.0015)	0.049 (0.0020)	0.053 (0.0012)	0.055 (0.0018)	0.057 (0.0018)	0.038 (0.0008)	0.036 (0.0007)	**0.0469**
SSC+BD+CEC	0.038 (0.0014)	0.038 (0.0014)	0.049 (0.0020)	0.051 (0.0012)	0.050 (0.0015)	0.050 (0.0016)	0.039 (0.0009)	0.036 (0.0007)	**0.0439**
SSC+BD+pH+CEC	0.045 (0.0017)	0.042 (0.0016)	0.049 (0.0020)	0.053 (0.0012)	0.055 (0.0018)	0.056 (0.0018)	0.038 (0.0007)	0.036 (0.0007)	**0.0468**
SSC+OC+pH+CEC	0.086 (0.0021)	0.077 (0.0019)	0.058 (0.0020)	0.061 (0.0018)	0.063 (0.0023)	0.062 (0.0022)	0.036 (0.0011)	0.035 (0.0012)	**0.0598**
SSC+BD+OC+pH+CEC	0.052 (0.0025)	0.047 (0.0022)	0.051 (0.0023)	0.054 (0.0015)	0.056 (0.0019)	0.056 (0.0019)	0.035 (0.0009)	0.032 (0.0009)	**0.0479**

	R^2								**AvgR^2**
SSC	0.315 (0.0225)	0.400 (0.0216)	0.604 (0.0144)	0.844 (0.0094)	0.888 (0.0086)	0.894 (0.0079)	0.910 (0.0040)	0.910 (0.0039)	**0.7206**
SSC+BD	0.661 (0.0140)	0.654 (0.0114)	0.654 (0.0089)	0.851 (0.0057)	0.891 (0.0058)	0.893 (0.0056)	0.908 (0.0038)	0.911 (0.0033)	**0.8029**
SSC+OC	0.418 (0.0244)	0.496 (0.0236)	0.652 (0.0123)	0.832 (0.0114)	0.879 (0.0112)	0.889 (0.0099)	0.919 (0.0058)	0.917 (0.0066)	**0.7503**
SSC+BD+OC	0.624 (0.0185)	0.641 (0.0156)	0.667 (0.0090)	0.851 (0.0065)	0.896 (0.0057)	0.901 (0.0054)	0.921 (0.0043)	0.920 (0.0046)	**0.8026**
SSC+pH	0.280 (0.0201)	0.412 (0.0192)	0.656 (0.0129)	0.791 (0.0160)	0.841 (0.0146)	0.853 (0.0132)	0.888 (0.0041)	0.887 (0.0042)	**0.7010**
SSC+CEC	0.324 (0.0256)	0.407 (0.0241)	0.619 (0.0123)	0.838 (0.0098)	0.877 (0.0095)	0.883 (0.0087)	0.895 (0.0050)	0.898 (0.0049)	**0.7176**
SSC+pH+CEC	0.311 (0.0202)	0.445 (0.0187)	0.682 (0.0120)	0.812 (0.0109)	0.860 (0.0093)	0.869 (0.0084)	0.886 (0.0045)	0.886 (0.0050)	**0.7189**
SSC+OC+pH	0.337 (0.0230)	0.452 (0.0200)	0.669 (0.0110)	0.802 (0.0103)	0.859 (0.0088)	0.872 (0.0082)	0.900 (0.0064)	0.898 (0.0077)	**0.7236**
SSC+OC+CEC	0.404 (0.0233)	0.479 (0.0205)	0.652 (0.0100)	0.827 (0.0113)	0.872 (0.0113)	0.884 (0.0100)	0.917 (0.0050)	0.917 (0.0059)	**0.7440**
SSC+BD+pH	0.590 (0.0128)	0.644 (0.0105)	0.666 (0.0116)	0.821 (0.0066)	0.863 (0.0062)	0.868 (0.0060)	0.892 (0.0043)	0.898 (0.0038)	**0.7803**
SSC+BD+CEC	0.676 (0.0131)	0.672 (0.0101)	0.656 (0.0090)	0.841 (0.0067)	0.878 (0.0067)	0.881 (0.0066)	0.902 (0.0043)	0.908 (0.0034)	**0.8018**
SSC+BD+pH+CEC	0.591 (0.0126)	0.649 (0.0102)	0.670 (0.0113)	0.821 (0.0065)	0.863 (0.0061)	0.867 (0.0057)	0.892 (0.0041)	0.899 (0.0036)	**0.7815**
SSC+OC+pH+CEC	0.346 (0.0239)	0.461 (0.0201)	0.676 (0.0103)	0.807 (0.0098)	0.864 (0.0084)	0.876 (0.0076)	0.901 (0.0059)	0.899 (0.0071)	**0.7288**
SSC+BD+OC+pH+CEC	0.574 (0.0174)	0.636 (0.0156)	0.673 (0.0104)	0.824 (0.0076)	0.872 (0.0069)	0.879 (0.0069)	0.910 (0.0049)	0.915 (0.0047)	**0.7854**

† Standard deviations of MD, RMSD and R^2 values generated by ensemble of k-NN estimations based on 100 replicates are presented in brackets. SSC is sand (%), silt (%) and clay (%), BD is bulk density (Mg m^{-3}), OC is organic carbon (%), pH is potential Hydrogen (-), CEC is cation exchange capacity (cmol kg^{-1} soil).

Table 4. Prediction performance in terms of RMSD of the k-NN method using four combinations of input attributes, of the PTFs of Hodnett and Tomasella (2002) and Minasny and Hartemink (2011).†

	RMSD (m^3 m^{-3})							
PTFs	θ_{0kPa}	θ_{-1kPa}	θ_{-3kPa}	θ_{-10kPa}	θ_{-20kPa}	θ_{-50kPa}	$\theta_{-250kPa}$	$\theta_{-1500kPa}$
k-NN (SSC+BD)	0.039 (0.0014)	0.039 (0.0014)	0.050 (0.0020)	0.051 (0.0013)	0.050 (0.0016)	0.051 (0.0017)	0.039 (0.0010)	0.036 (0.0008)
k-NN (SSC+OC)	0.078 (0.0022)	0.070 (0.0021)	0.057 (0.0021)	0.058 (0.0018)	0.057 (0.0024)	0.056 (0.0023)	0.034 (0.0012)	0.033 (0.0013)
k-NN (SSC+BD+CEC)	0.038 (0.0014)	0.038 (0.0014)	0.049 (0.0020)	0.051 (0.0012)	0.050 (0.0015)	0.050 (0.0016)	0.039 (0.0009)	0.036 (0.0007)
k-NN (SSC+BD+OC+pH+CEC)	0.052 (0.0025)	0.047 (0.0022)	0.051 (0.0023)	0.054 (0.0015)	0.056 (0.0019)	0.056 (0.0019)	0.035 (0.0009)	0.032 (0.0009)
Hodnett and Tomasella (2002)	0.036 (-)	0.042 (-)	0.059 (-)	0.049 (-)	0.046 (-)	0.041 (-)	0.036 (-)	0.035 (-)
Minasny and Hartemink (2011)	-	-	-	0.062 (-)	-	-	-	0.045 (-)

† Standard deviations of RMSD values generated by ensemble of k-NN estimations based on 100 replicates are presented in brackets. SSC is sand (%), silt (%) and clay (%), BD is bulk density (Mg m^{-3}), OC is organic carbon (%), pH is potential Hydrogen (-), CEC is cation exchange capacity (cmol kg^{-1} soil).