Published online November 20, 2006

Sensitivity Analysis of the Nonparametric Nearest Neighbor Technique to Estimate Soil Water Retention

A. Nemes,* W. J. Rawls, Ya. A. Pachepsky, and M. Th. van Genuchten

ABSTRACT

A *k*-nearest neighbor (*k*-NN) nonparametric algorithm variant was earlier applied successfully to estimate soil water retention. In this study, we tested the sensitivity of that *k*-NN variant to different data and algorithm options, such as: (i) estimations made to soils with differing distribution of properties; (ii) the use of different sample weighting methods; (iii) the number of ensembles we developed; (iv) data density in the reference data set; (v) the presence of outliers in the reference data set; (vi) unequal weighting of input attributes; and (vii) the addition of locally specific data to the reference data set. We used a hierarchical set of input attributes and data set sizes to develop ensembles of predictions using multiple randomized subset selections. The *k*-NN technique performed comparably well as neural network models developed on the same data. Using >50 ensemble members did not improve the results any further. The *k*-NN technique showed little sensitivity to the choice of sample weighting methods and to suboptimal weighting of input attributes. Differences in data density in parts of the reference data set did not substantially impact estimation errors. Estimations substantially improved for locally specific data when some local samples were included in the reference data set, while estimations for other samples remained almost unaffected. The *k*-NN technique shows a large degree of stability and insensitivity to different settings and options, can easily adopt new data without the need to redevelop equations, and is an effective alternative to other techniques to estimate soil water retention.

Modeling water and solute transport has become an important part of simulating agricultural productivity as well as environmental quality. The use of models, however, is often hindered by the lack of information on soil hydraulic properties. For many applications, the estimation of those properties using pedotransfer functions (PTFs) is a feasible alternative to costly and time-consuming measurements.

One common feature of today's PTFs is that they are all based on some parametric approach, i.e., they are equations with parameters found from fitting those equations to data. Identifying the right equation and ensuring that the associated probability distributions of errors will be similar across the data space is not always easy. Estimation results can be heavily biased in the case of small sample size in the development data set. The equations need to be redeveloped and republished, should new data become available, and users are not able to simply include any additional data sets to improve performance for their site-specific range of soil properties.

An alternative approach for such estimations is the use of nonparametric techniques. Such techniques are based on pattern recognition rather than on fitting equations to data. One of these techniques is the (*k*-)nearest neighbor (*k*-NN), which has been widely applied in pattern recognition and statistical classification tasks (Dasarathy, 1991). This technique belongs to the group of "lazy learning algorithms". It is "lazy" in that it passively stores the data until the time of application; all calculations are performed only when estimations need to be generated. Applications of this technique can be found in the literature of many fields; some recent applications in the fields of meteorology and hydrology are Yakowitz (1993), Lall and Sharma (1996), Tarboton et al. (1998), Sharma and O'Neill (2002), Harrold et al. (2003a, 2003b), and Mehrotra and Sharma (2006).

In a soil physical and hydrophysical context, a similarity-based *k*-NN type technique has been applied successfully by Nemes et al. (1999) to interpolate soil particle-size distributions. They found this technique to perform well while estimating the missing 50-μm particle fraction for many European soils, which later served as input to soil hydraulic PTFs. Jagtap et al. (2004) introduced a dynamic *k*-NN technique to estimate the drained upper limit and lower limit of plant water availability from field-measured soil water retention information. They compared their model to existing soil hydraulic PTFs to make estimations for their data set and concluded that the *k*-NN technique performs better than three published regression-type parametric PTFs. Most recently, Nemes et al. (2006) estimated soil water content at −33 and −1500 kPa matric potentials using a *k*-NN variant and a hierarchical set of input data that originated from the USA. In this study, the *k*-NN technique was found to be competitive with neural network (NNet) models that are cited as probably the most advanced and accurate PTF technique of the day.

Application of the k-NN technique means identifying and retrieving the nearest (most similar) stored objects to the target object. The quality of such estimations depends on, among others, which objects are considered to be the nearest to the target object. An understanding of how this technique works suggests, however, that

A. Nemes, Univ. of California, Dep. of Environmental Sciences, Riverside, CA 92521; A. Nemes and W.J. Rawls, USDA-ARS Hydrology and Remote Sensing Lab., 10300 Baltimore Ave., Bldg. 007, BARC-West, Beltsville, MD 20705; Ya.A. Pachepsky, USDA-ARS Environmental Microbial Safety Lab., Powder Mill Road, Bldg. 173, BARC-East, Beltsville, MD 20705; and M.Th. Van Genuchten, USDA-ARS George E. Brown Jr. Salinity Lab., 450 West Big Springs Rd., Riverside, CA 92501. Received 25 Jan. 2006. *Corresponding author (anemes@hydrolab.arsusda.gov).

Published in Vadose Zone Journal 5:1222–1235 (2006).
Original Research
doi:10.2136/vzj2006.0017

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: OM, organic matter; SSC, sand, silt, and clay contents (soil texture); NNet, neural network; *k*-NN, *k*-nearest neighbor technique; PTF, pedotransfer function; RMSR, root mean squared residual; θ33, water retention at −33kPa matric potential; θ1500, water retention at −1500 kPa matric potential.

there are several factors that may potentially have a great influence on which objects are ruled to be nearest, and how their influence on the final estimate is accounted for. A standard *k*-NN does not perform attribute selection; it allows irrelevant or interacting inputs to have as much effect on the distance calculation as any other useful inputs. Some inputs may also have a (considerably) wider numerical range of data than others. A unit change in one input variable may have a much larger influence on the distance measure than the same change in the other. Such concerns lead to the introduction of data normalization systems and different attribute weighting systems in more recent *k*-NN variants (e.g., Wettschereck et al., 1997; Mehrotra and Sharma, 2006). Different recommendations exist for the weighting of the retrieved nearest objects while formulating the output of the *k*-NN technique. Simple averaging of the output attributes of the retrieved neighbors is probably the simplest solution, but it does not consider any difference in the resemblance of the retrieved neighbors to the target object. Lall and Sharma (1996) recommended a method that assigns weights based on the rank of each of the retrieved *k* neighbors in being the nearest to the target object. Yates et al. (2003) derived realizations of the output variable by the selection of a single neighbor from the retrieved *k* neighbors using a random number generator. Nemes et al. (2006) used a weighting method that assigns weights based on the distances of the retrieved *k* neighbors from the target object. Such methods have not previously been tested against each other while being applied to estimate unsaturated soil hydraulic properties.

It is unknown to what extent estimations are sensitive to local data density in the underlying development (reference) data set. In other words, it has not been quantified how reliable the estimates are for a type of soil that can be found in the reference data set but is relatively poorly represented. The above weighting methods may perform differently, and the estimations may be biased or generally be less accurate for cases that are poorly represented in the reference data set. Pedotransfer functions are expected to give reliable estimates for soils that are from the same population as the data set used to develop the PTF; however, PTFs are frequently sought and tested that work well for soils that originate from a population different than that used to develop the PTF. Schaap and Leij (1998) showed the great extent to which PTFs could be dependent on the origin (and distribution) of the soils in the development and application data sets. In the study of Nemes et al. (2006), the *k*-NN technique was tested using soils that were not in the reference data set, but that originated from the same data source and distribution.

The approach of using an ensemble of estimation or forecast techniques is widely used in meteorology (e.g., Molteni et al., 1996; Houtemaker et al., 1996; Palmer et al., 2004). The approach essentially means averaging the predictions of a number of models that are applied to the data simultaneously. The rationale behind such approach is that the use of a particular single model is often not justifiable. Using multiple models, features that are consistent among ensemble members will be preserved through averaging, while those that are inconsistent will be reduced in amplitude. In the meantime, the output of each member can be viewed as a potential sample from the outcome of the estimations and be used to calculate estimation uncertainty. In subsurface hydrology, Ye et al. (2004) suggested averaging of the spatial variability models in unsaturated fractured tuff. Guber et al. (2006) tested the performance of an ensemble of 22 PTFs to estimate soil water retention against measured data and subsequently used such data in soil water flow simulations with success.

An alternative form of ensemble predictions is when one uses multiple realizations of the PTF development data set, obtained by, for example, randomized subset selection or bootstrapping. In this way, multiple subestimates are obtained that can be interpreted as samples from the statistical distribution of estimation outcomes and can subsequently be used to characterize estimation uncertainty. Examples of such estimations can be found in, e.g., Schaap and Leij (1998), Schaap et al. (1999), Nemes et al. (2003, 2006), and Baker (2005). It is not known, however, how many ensembles are minimally needed to obtain estimates that are not significantly changed by the addition of an additional ensemble member. Generating an excessive number of ensembles may not be beneficial in terms of improving the estimations, while it increases computation time.

The objective of this study was to test the sensitivity of the *k*-NN variant introduced by Nemes et al. (2006) to different algorithm options and to differences in the properties of the underlying data. We also compared the performance of the *k*-NN technique to the performance of NNet models developed using the same data.

MATERIALS AND METHODS

Soil Data

This work comprises seven case studies. In most case studies, we used data from two data sets described below. In Case Study 7 we used an additional third data set. The first dataset encompasses 2125 soil horizons that were selected from the NRCS Soil Characterization Database (Soil Survey Staff, 1997), according to the following criteria: (i) mineral soil horizons were selected from the contiguous USA having horizon notation A, A1, and Ap (and their derivatives), with the condition that the top of the horizon was at the soil surface; (ii) organic matter (OM) content of the selected soils was limited to 1 to 15%, and (iii) bulk density (D_b) was limited to 0.5 to 2.0 g cm^{-3}. Selected soil properties were the following: sand (50–2000 μm), silt (2–50 μm), and clay content (<2 μm) according to the USDA classification system (Soil Survey Staff, 1951), D_b, OM content, and retained (volumetric) water at −33 and −1500 kPa matric potentials, (θ33 and θ1500, respectively), with no missing data allowed in any of the fields. Such matric potentials were chosen as those are often used to approximate field capacity and the wilting point when calculating plant-available water, and thus are often preferred points in water retention curve (WRC) determinations in the laboratory. Measured WRC data at those matric potentials can be found frequently in many soil hydraulic databases worldwide. No entries were allowed that showed obvious inconsistency in physical or hydraulic data (sand + silt + clay ≠ 1; θ33 <

θ1500; [(1 − [D_b]/2.65) − θ33] < 0). This data set is referred to below as NRCS, and has been used as reference data as well as to provide data for testing the estimations. Samples have been randomly drawn to be either a member of the reference data set or a test data set. We elected to use 435 samples, i.e., ~20% of all data, as test data in each case. We used different sizes of reference data sets to evaluate the effect of the size of the reference data set on each examined factor. Samples were drawn to be members of reference data sets of 100, 200, 400, 800, and 1600 samples. All random data selections were repeated 200 times to allow the development of an ensemble of PTF estimations. By using a sufficiently large number of ensembles, the impact on the final estimation results of any single ensemble (i.e., any particular data set division) can be minimized. Optimization of the number of ensembles was part of this study.

A second data set from the European HYPRES database (Wösten et al., 1999) served as an alternative test data set in most of the case studies. These data originated from a different geographical area and show a different data distribution than the NRCS data set. The HYPRES data were donated by researchers of some 20 institutions of different countries in Europe. The selection process and criteria were the same as for the NRCS data set, with one exception: there are differences in the water retention data reported by the different sources. We allowed water retention data reported in the ranges of −30 to −34 kPa and −1500 to −1600 kPa matric potentials to represent θ33 and θ1500, respectively. Altogether, we selected 435 samples from HYPRES that were used solely as an independent second test data set, referred to as HYPRES.

An additional third data set was used in Case Study 7 below. This data collection originated from Brazil and has been used before by, e.g., Tomasella et al. (2000, 2003). To avoid the presence of outliers, the same limitations as for the NRCS data set were imposed on the source data collection. This yielded a set of 428 samples, which are referred to as BRAZ.

Table 1 shows the summary statistics of selected soil attributes of the selected data sets. The data sets contain data on a wide range of soils, in terms of the shown soil attributes. The average sand content of the HYPRES soils is about 10% larger than that of the NRCS soils, whereas its silt content is less by close to the same amount. The NRCS data set represents a substantially wider range of soils in terms of silt and clay content than the HYPRES data set. Differences in other listed properties are less noticeable. We converted gravimetric water contents stored in the NRCS database to volumetric water contents to remain compatible with most existing PTFs and between data sets that we used. Different D_b values are stored in the NRCS database—measured at different states of wetness—that had to be used to convert −33 kPa and −1500 kPa gravimetric water contents to their respective volumetric water content values. While the NRCS and HYPRES data sets consist of soils that originated almost exclusively from areas with temperate (continental or maritime) and Mediterranean climates, most soils in the BRAZ data set originated from areas with a tropical climate. A known physical characteristic of many soils in the tropics is the bimodality of their particle-size distribution (MacLean and Yager, 1972; Tomasella et al., 2003). This results in a composition of the solid phase that is rich in sand and clay particles while, in many soils, the silt fraction is almost completely missing. This characteristic can be identified in the BRAZ data set in Table 1; this set has, on average, a substantially larger proportion of clay and lower proportion of silt contents than the other two data sets. This data set also has a considerably lower average D_b than the other two sets.

Table 1. Summary statistics of selected soil attributes† in the data sets.

	Sand	Silt	Clay	D_b	OM	θ33	θ1500
	kg kg^{-1}			g cm^{-3}	%	m^3 m^{-3}	
NRCS							
Min.	0.004	0.034	0.002	0.520	1.000	0.051	0.022
Max.	0.955	0.922	0.811	1.890	14.861	0.724	0.725
Mean	0.280	0.492	0.228	1.362	3.082	0.316	0.171
SD	0.231	0.194	0.133	0.186	2.063	0.083	0.094
Median	0.211	0.491	0.205	1.380	2.500	0.325	0.153
HYPRES							
Min.	0.007	0.040	0.023	0.899	1.000	0.047	0.039
Max.	0.931	0.791	0.670	1.700	13.700	0.583	0.422
Mean	0.383	0.407	0.210	1.402	2.688	0.292	0.170
SD	0.289	0.212	0.138	0.170	1.944	0.105	0.090
Median	0.291	0.397	0.188	1.430	2.090	0.298	0.139
BRAZ							
Min.	0.000	0.034	0.040	0.720	1.000	0.080	0.023
Max.	0.910	0.710	0.810	1.760	11.016	0.548	0.414
Mean	0.271	0.230	0.499	1.173	2.730	0.321	0.229
SD	0.225	0.165	0.214	0.204	1.532	0.093	0.080
Median	0.180	0.190	0.528	1.176	2.300	0.333	0.238

† D_b = bulk density; OM = organic matter content; θ33 = water retention at −33 kPa matric potential; θ1500 = water retention at −1500 kPa matric potential.

The *k*-Nearest Neighbor Technique

Unlike classic PTFs, the *k*-NN technique does not use any predefined mathematical functions to estimate a certain attribute. A "reference" data set—analogous to the development or training data sets used to develop classic PTFs—is searched for samples that are most similar to the target sample, based on selected input attributes. In most classic *k*-NN variants, the "distance" measure is calculated as the classical Euclidean distance between the target and the known instances (Wettschereck et al., 1997). In a simple case with only two input attributes, e.g., sand and clay content, selection of the nearest (or most similar) soil(s) can be represented geometrically using Pythagoras' theorem, as demonstrated by, e.g., Jagtap et al. (2004). The "distance" of each soil from the target soil can be calculated as the square root of the sum of squared differences in sand and clay content between the target soil and each of the soils of the reference data set. Soils of the reference data set will then be sorted in ascending order of their distance from the target soil. The estimated value of the output attribute is calculated as the weighted average of the output attribute of a preselected number of the nearest soils.

Most PTFs, however, use information on more than two input attributes. For such cases, the generalized form of

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

provides a sufficient solution, where d_i is the "distance" of the ith soil from the target soil, Δa_{ij} represents the difference of the ith soil from the target soil in the jth soil attribute, and J is the total number of soil attributes considered as inputs. The term *distance* does not refer to actual (physical) distance, but to a measure of similarity; the distance will be smaller for soils that are more similar to the target soil in their input attributes.

A unit difference in one attribute may, however, not be as influential as the same unit difference in another attribute. For example, sand content, if given as a percentage, can take values anywhere between 0 and 100, whereas D_b content ranges theoretically from 0 to a maximum of 2.65 g cm^{-3} in soils. In

real data sets, the range of such values is usually narrower. A unit difference in D_b is expected to be more influential than the same unit difference in sand content. To avoid bias toward one attribute or the other, the data need to be normalized before they are used to calculate "distance" using Eq. [1].

Nemes et al. (2006) suggested the following normalization. First, all input attributes are transformed to obtain temporary variables with distribution having zero mean and a standard deviation of 1:

$$a_{ij(\text{temp})} = [(a_{ij}) - \overline{a}_j] / \sigma(a_j) \quad [2]$$

where $a_{ij(\text{temp})}$ represents the temporary value of the jth attribute of the ith soil, and $\overline{a}_j$ and $\sigma(a_j)$ represent the mean and standard deviation of the observed values of the jth attribute in the reference data set. The difference between the minimum and maximum of those temporary variables is then examined, and the attribute that shows the widest range of transformed (temporary) values is identified. The ratio of the widest range of transformed values and the range of transformed values of each attribute were used as a scaling factor to obtain zero mean and the same minimum–maximum range in the data of all attributes:

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

where $a_{j(\text{temp})}$ represents the temporary data of the jth soil attributes normalized using Eq. [2], and $a_{ij(\text{trans})}$ represents the final transformed values of the jth attribute of the ith soil that are to be used as input.

Nemes et al. (2006) experimented with the number of soils used to obtain the estimate of the output attribute of the target soil (k). They suggested using $k = 0.655N^{0.493}$ to calculate the optimal value of k based on the size of the reference data (N) set. Alternatives exist to this method. Lall and Sharma (1996), for instance, suggested $k = N^{1/2}$ for $N > 100$, based on their experience under certain conditions, where N is the length of the observed sample record, i.e., the number of known instances in the reference data set. Both studies note, however, the relatively low sensitivity of the technique to the choice of k. In this study, we applied the formula suggested by Nemes et al. (2006).

The user of the k-NN technique also has to decide how to weight each selected soil while calculating the estimate of the output attribute. As one solution, the simple average of their output attribute can be calculated. In such case, the weight $[w_{(i)}]$ of each selected soil will equal $1/k$, where k is as defined above. The calculated "distance" of each soil from the target object (see Eq. [1]) will be different, however, and it can be argued that a soil closer to the target object should have more weight in calculating the estimated value than a soil that is further from it. Weighting methods that allow either rank- or distance-dependent weighting of soils offer alternative solutions. One solution mentioned in the literature is that of Lall and Sharma (1996), who calculated weights for each selected neighbor as

$$w_{(i)} = \frac{1/i}{\sum_{i=1}^{k} 1/i} \quad [4]$$

where $w_{(i)}$ is the weight associated with the ith nearest neighbor and k is the number of neighbors considered. This method considers the rank of each sample in being the nearest neighbor to the target object, and does not consider the relative distances of the selected k neighbors from the target object. Nemes et al. (2006) used a weighting method that accounts for the distribution of distances of the k nearest neighbors from the target object as follows:

$$w_i = d_{i(\text{rel})} / \sum_{i=1}^{k} d_{i(\text{rel})} \quad [5]$$

where k is the number of nearest neighbors retrieved, w_i is the assigned weight, and $d_{i(\text{rel})}$ is the relative distance of the ith nearest neighbor, calculated as

$$d_{i(\text{rel})} = \left(\sum_{i=1}^{k} d_i / d_i\right)^p \quad [6]$$

where k is the number of neighbors considered, d_i is the distance of the ith selected neighbor calculated using Eq. [1], and p is a power term that was optimized to provide the best estimation results. The p term accounts for the weight/distance relationship and was suggested to be set at $p = 0.767N^{0.049}$, where N is the number of samples in the reference data set.

Nemes et al. (2006) used different (hierarchical) combinations of the following input attributes: USDA sand, silt, and clay content (SSC), D_b, and OM content. They assumed that these attributes are all equally relevant and important in the estimation of the output attributes. Four different sets of input attributes were used to estimate θ33 and θ1500 from data of the NRCS data set. The simplest model used only SSC as predictors. In the following two models, either D_b, or OM content was added to SSC as a predictor (SSCBD and SSCOM, respectively). In the fourth model, all of these inputs were used as predictors (SSCBDOM). In this study, we implemented this approach to avoid a possible bias while applying one particular set of input attributes, and to account for different levels of data availability for potential future users.

The Artificial Neural Network Technique

Recently, artificial NNet models have been used successfully in PTF development (e.g., Pachepsky et al., 1996; Tamari et al., 1996; Schaap et al., 1998; Koekkoek and Booltink, 1999; Minasny et al., 1999; Schaap and Leij, 2000; Minasny and McBratney, 2002; Nemes et al., 2003). Most studies found that the predictive capabilities of NNet PTFs were equivalent or superior to different regression-type PTFs. For this reason, we chose the NNet technique to serve as the basis of comparison for the k-NN technique in Case Study 1.

A NNet model consists of many simple computing elements (termed *neurons* or *nodes*), that are organized into subgroups (layers) and are interconnected as a network by weights. A model typically consists of an input layer, an output layer, and one (or more) "hidden" layer(s) that connect(s) the input and output layers. The number of nodes in the input and output layers correspond to the number of input and output variables of the model; the number of hidden nodes can be varied freely. Data flow goes from the input layer through the hidden layer(s) to the output layer. A node in the hidden and output layers receives multiple inputs—typically from all nodes of the previous layer. Within the node, each input is weighted and combined to produce a single value as the output of that node, which is then directed to all the nodes of the next layer, or outputted if it was a node of the output layer. The weight matrices are obtained through a calibration (training) procedure, which can then be used to make estimations for independent data. For a more thorough description on NNets, see Hecht-Nielsen (1990) or Haykin (1994).

Following Nemes et al. (2006), we used a three-layer backpropagation NNet model. There are different approaches to set the number of nodes in the hidden layer. We elected to calculate it as half of the total number of input and output variables, rounded up to the nearest integer. Four different models, each using a different set of input attributes, were developed to estimate θ33 and θ1500 separately from data of

the NRCS data set. To allow direct comparison with the performance of the *k*-NN models, inputs to the four NNet models were the same as inputs to the four *k*-NN models outlined above. We transformed all data, before being presented to the NNets, to take up the interval [0,1].

The NNets were combined with the data selection procedure of the bootstrap method (Efron and Tibshirani, 1993) to generate internal calibration–validation data set pairs for an early stopping procedure. We generated 10 bootstrap replica data sets, each of which was used to calibrate the NNet models. This procedure provided 10 subestimates that could be slightly different from each other. The estimate of a PTF from one particular ensemble data set—for each single value—was then calculated by averaging the 10 subestimates of the value. Application of the bootstrap method took place internally in the NNet program to derive the best estimates from each ensemble's development data set, and was performed independently within each of the 200 PTF ensembles described above. All NNet modeling was performed using the Neural Network Toolbox in MATLAB (Demuth and Beale, 1992).

Evaluation Criteria

In Case Study 1, *k*-NN estimations were compared with NNet estimations. Other case studies provided comparisons made only between *k*-NN models using different settings. In all cases, the goal was to report on or to minimize the estimation errors for the test data set(s) at the population level, which was characterized by two measures. Root mean squared residual (RMSR) of the estimations is defined as

$$\mathrm{RMSR} = \sqrt{(1/N)\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i)^2} \qquad [7]$$

and the mean residual value (MR) is calculated as

$$\mathrm{MR} = (1/N)\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i) \qquad [8]$$

In Eq. [7] and [8], N is the number of samples in the test data set, θ and $\hat{\theta}$ are measured and estimated water contents, respectively. The MR can quantify systematic errors between measurements and estimations and the RMSR can give the accuracy of the estimations in terms of standard deviations.

For Case Study 4, however, we do not directly report on MR or RMSR values, as the actual values have no particular importance. Rather, we examined the correlation between the estimation error (residual) for the individual test soils and the distance of the retrieved neighbors from the target object, using R^2, the coefficient of determination, calculated as

$$R^2 = \frac{[N(\sum xy) - (\sum x)(\sum y)]^2}{\{N\sum[x^2 - (\sum x)^2]\}\{N\sum[\,y^2 - (\sum y)^2]\}} \qquad [9]$$

where x and y represent the independent and dependent variables in the equation, respectively, and N is the number of samples. The R^2 shows what proportion of the dependent variable can be attributed to the effect(s) of independent variable(s).

CASE STUDIES

Application of the *k*-Nearest Neighbor Technique to Data from an Alternative Data Source

The *k*-NN technique was applied to estimate θ33 and θ1500 for the HYPRES data set, using the NRCS data set as the reference data set. The rationale behind this case study was that PTFs usually produce worse estimations for data that originate from a different geographical area than for the data used to develop or train the PTF. We compared the degree of such loss in accuracy by the *k*-NN technique and an alternative technique. The *k*-NN technique was applied using each of five reference data set sizes (N = 100, 200, 400, 800, and 1600) and four input attribute sets (SSC, SSCBD, SSCOM, and SSCBDOM) to estimate two output attributes (θ33 and θ1500). To demonstrate the capabilities of this technique, NNet models were also applied to estimate the same output attributes, using the same sets and the same input attributes as for the *k*-NN technique. This way we show the pure difference between the two techniques, without being affected by differences originating from the underlying data.

Comparison of Methods to Weight the Retrieved *k* Neighbors

The user of the *k*-NN technique has to make a choice how to weight the retained neighbors while forming the estimate of the output attribute. We examined three alternative methods: (i) simple averaging of the output attributes of the retrieved neighbors; (ii) "rank"-based weighting according to Lall and Sharma (1996) (Eq. [4]); and (iii) distance-based weighting according to Nemes et al. (2006) (Eq. [5] and [6]). These weighting methods were implemented in the algorithm separately, each combined with each of the reference data set sizes, input attribute sets and output attributes, as outlined for Case Study 1. The NRCS data set was used as the reference data set, and estimations were made for soils of both the NRCS and HYPRES test data sets.

Sensitivity of the Estimation Accuracy to the Number of Model Ensembles

When multiple realizations are developed from the same master data set, usually the number of ensembles is set arbitrarily. It is rarely examined—and thus remains a question—whether the outcome of the estimations would be significantly changed if the developer chose to use more or fewer ensembles. This may result in suboptimal estimation result—when the number chosen is too small—or in unnecessarily long computations—when the number chosen is too large. We examined the effect of the number of ensembles using each of the five reference data set sizes, four input attribute sets, and two output attributes as outlined for Case Study 1. The NRCS data set was used as the reference data set, and estimations were made for soils in both the NRCS and HYPRES test data sets. Two hundred ensembles were run and running RMSRs were recorded after the completion of estimations using each ensemble.

Sensitivity of the *k*-Nearest Neighbor Technique to Data Density

One of the reported advantages of the *k*-NN technique is that, unlike parametric PTFs, it uses information

that is specific to the target object. It does so as the "nearest neighbors" are selected in terms of their properties, meaning that they are similar to the target object. Parametric PTFs do not work this way; one (set of) parametric equation(s) describes the entire data space. It can be argued, however, that estimation errors may be much larger when the selected "nearest" neighbors are still distant from the target object, i.e., when the data density in the reference data set is small at some locations in the data domain, meaning that some of the target objects are not well represented. We hypothesized that larger distances calculated according to Eq. [1] will lead to larger estimation errors. To test this, we correlated estimation errors with the calculated distances between the retrieved nearest neighbors and the target objects.

Sensitivity of the *k*-Nearest Neighbor Technique to the Potential Presence of Outliers

Input data are normalized before being presented to the calculation algorithm. This is done to assure that the different input attributes will receive equal weight in the distance calculations (Eq. [1]). It may happen, however, that the *k*-NN technique there are just a few outliers in the unique data set, which may expand the data range of one or more of the input attributes. When an outlier is present, it may not add much to the characterization of the whole data set, but may mask the "true" effective data range in the data set. The significance of this phenomenon in the proposed *k*-NN variant is that while performing the data normalization (i.e., Eq. [2] and [3]), an input attribute with outlier(s) may eventually get lower weight in the distance calculations than it would based on its effective data range. We examined the effect of the presence of outliers in the reference and test data sets by simulating the presence of outliers. In this case study, we used only the NRCS data set that was also used in the preceding case studies. Specifically for this case study, we imposed further limitations on this data set. We left out the samples with their sand content <0.1 or >0.8 kg kg^{-1}, and samples with OM content $<2\%$ or $>14\%$. As a result, we obtained a set of 650 samples. We then put only two of those discarded samples back in the data set in a controlled manner. We recorded whether those samples appeared (i) only in the reference data set, (ii) randomly in the reference or test data sets, or (iii) only in the test data set. We also ran a fourth option, in which neither appeared in any of the data sets. To stay consistent with the established reference data set size options, we elected to use 400 samples as a reference data set in each of the ensembles. All other samples were used as test data.

Unequal Input Attribute Weighting

It can be argued that equal weights assigned to each input attribute—as introduced by applying Eq. [2] and [3]—may not provide the best possible results, and that particular attribute(s) should receive more weight. As a simple example, we refer to the known dominance of clay content over that of sand content in the characterization of the dry end of the water retention curve, e.g., $\theta 1500$ in our study. This is because clay particles have a significantly larger role in determining the presence and distribution of finer pores in the soil, which control soil hydraulic properties in the dry range. To test the performance and sensitivity of the *k*-NN technique to unequal attribute weighting, we introduced additional scaling to each of the input attributes after normalizing them first according to Eq. [2] and [3]. This concept has the same logic as the influence weight concept of Mehrotra and Sharma (2006). By applying all combinations of weight factors of 1, 5, 10, 30, 100, and 200, we introduced scaling to cover ratios of 1, 2, 3, 3.33, 5, 6, 6.66, 10, 20, 30, 40, 100, and 200 to 1. All combinations of all the above ratios were applied to all input attribute combinations in the algorithm that used all listed input attributes (SSCBDOM) and the reference data set size of 1600 samples. We used the NRCS test data set to test the performance of each of the weight combinations. Root mean squared residual values were ordered and combinations of attribute weights that yielded the smallest RMSRs were logged. Models were compared with each other and with the base model with 1:1 attribute weights.

Estimations using Locally Specific Data

An advantage of this nonparametric technique is that it is capable of easily adopting new, locally specific data into a general reference data set that had been previously established. Should new data become available, the user is able to include those in the reference data set without the need to redevelop or republish any equations or calculation matrices. A user will presumably be able to improve estimations for specific local samples by incorporating existing local information in the reference data set, without affecting estimations for other sections of the data space. We tested this potential feature of the *k*-NN technique by introducing to the study an additional data set that originated from Brazil. This data collection originated primarily from areas with a tropical climate, as opposed to our previously used two data sets, NRCS and HYPRES.

It was hypothesized that the addition of "locally specific" samples would only have noticeable impact on estimations made to other samples from this location without causing significant alteration or degradation in the performance for other samples. This was assumed because the addition of locally specific data changes or improves data density locally in that specific part of the data space. Using the *k*-NN technique means that we selected samples from the close neighborhood of the target object—in terms of its properties—from the reference data set. For soils with other textures, the original reference data set (without locally specific data) would supposedly provide the same estimations as before, as the locally specific samples would not be selected as the nearest neighbors because of the larger differences in their properties. We used the default NRCS data set to provide samples for the reference data set, similar to the other case studies. We used 1600 samples. As an option, additional data from the BRAZ data set was also used as

input. Figure 1 shows the particle-size distribution of the soils in all three data sets. We divided the BRAZ data set into two parts. The section above the divider lines (labeled [OUT]) contains soils the texture of which are practically not found in the NRCS and HYPRES data sets. This is despite the fact that simple summary statistics did not reveal that (c.f., Table 1). The divider lines are defined by the CLAY(kg kg^{-1}) $>$ 0.7–0.75 $\times$ SAND(kg kg^{-1}) and CLAY(kg kg^{-1}) $>$ 0.1 equations. To account for the presence of locally specific data in the reference data set, the BRAZ[OUT] data were randomly split, and part of the data (108 in each case) were optionally added to the original reference data sets. The rest of the BRAZ[OUT] data set (N = 160) was used for testing, along with the NRCS test set, and HYPRES and BRAZ[IN] (N = 160) data sets. Ensemble estimations of both θ33 and θ1500 water contents were made using a reference data set of 1600 NRCS soils with the 108 BRAZ[OUT] soils optionally added. We used two different sets of input attributes: (i) textural properties only (SSC); and (ii) textural properties plus D_b and OM content (SSCBDOM).

RESULTS AND DISCUSSION

Application of the *k*-Nearest Neighbor Technique to Data from an Alternative Data Source

Using the design parameters k and p, calculated after Nemes et al. (2006), we applied the k-NN technique, using all five data set sizes and four input attribute sets, and made estimations for both output attributes for the HYPRES data set. We also performed the same estimations on identical data using NNet models. Results, in terms of RMSR, are summarized in Table 2. Root mean squared residual values are shown separately for each output attribute, estimation technique, input attribute set, and each development or reference data set size. Trends that can be observed in Table 2 correspond to those reported by Nemes et al. (2006): more accurate estimations of θ1500 than of θ33; a slight improvement in the estimations when more input attributes were used; and the mostly insignificant loss of estimation accuracy with a smaller number of samples used in the reference data set. These apply to both techniques. Root mean squared residuals for the HYPRES data set are, in general, 0.01 to 0.016 m^3 m^{-3} worse than those reported by Nemes et al. (2006) for the NRCS data set. This is observed using both techniques, and is consistent with the findings of, e.g., Schaap and Leij (1998). Estimations made for a data set having different data characteristics than the development data set is expected (and was shown) to be worse than for a data set having the same characteristics. This was also the case in this study, independent of which estimation technique was used. When the two techniques are compared pairwise, the NNet model resulted in 0.001 to 0.005 m^{3}m^{-3} smaller average RMSRs; such differences are only 0.001 m^3 m^{-3} larger than those reported by Nemes et al. (2006). Lesser estimation accuracy for the HYPRES data set does not seem to be due to lesser capabilities and suboptimal settings of the k-NN technique because the NNet model lost accuracy comparably.

Comparison of Methods to Weight the Retrieved *k* Neighbors

We compared three weighting methods that are different in their fundamentals. Simple averaging naturally means that for a given estimation, any of the selected k neighbors will carry equal weights. The method of Lall and Sharma (1996) assigns weight to each of the selected k neighbors based on their rank in similarity to the target object in their input attributes. The method of Nemes et al. (2006) accounted for the distribution of their actual degree of similarity to the target object. Results have been averaged by the input and output attributes, and are presented for each weighting method and reference data set size in Fig. 2. The differences between RMSRs at the population level using the three weighting methods are statistically insignificant within each test data set and reference data set size. This is somewhat surprising, because it suggests that no differentiation is really necessary among the k number of selected samples, regardless of their actual resemblance to the target object. This may still be understandable, given their closeness; the selected k samples all have significant relevance to the target object. There are special cases, however, when the target sample falls close to the edge of the data domain in one or more properties, or the combination of its properties may simply not be well represented in the reference data set. In these cases, some of the selected k neighbors may fall a lot farther from the target sample in their properties than other selected samples. Neighbors that are significantly farther than others should theoretically get much less

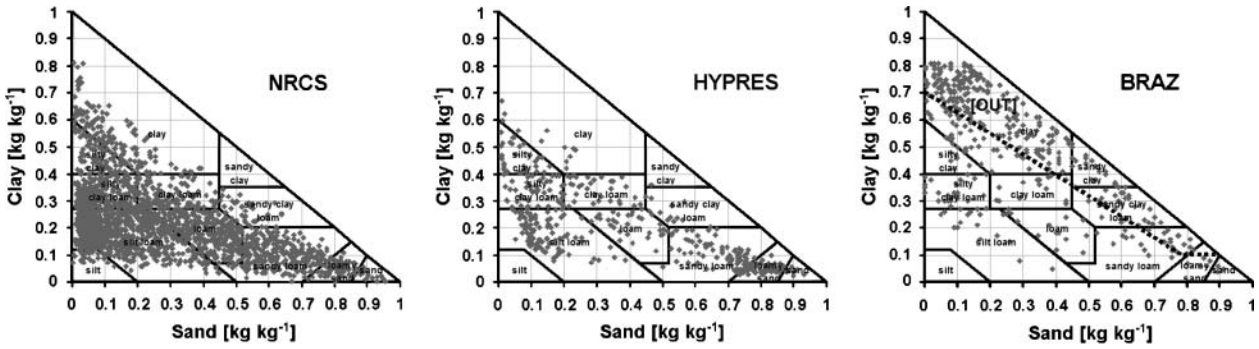

Fig. 1. Distribution of samples in the NRCS, HYPRES, and BRAZ data sets according to the NRCS textural triangle.

Table 2. Root mean squared residuals for the HYPRES data set using the *k*-nearest neighbor technique with settings according to Nemes et al. (2006), and the neural network models.

			Sample size of the pedotransfer function development data set (*N*)									
			N = 1600		*N* = 800		*N* = 400		*N* = 200		*N* = 100	
Estimated attribute	Estimation method	Input attributes†	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Water retention at −33 kPa	Nearest neighbor	SSC	0.069	<0.001	0.069	0.001	0.070	0.001	0.071	0.002	0.072	0.003
		SSC, D_b	0.067	<0.001	0.067	0.001	0.068	0.001	0.068	0.002	0.070	0.003
		SSC, OM	0.066	<0.001	0.066	0.001	0.066	0.001	0.067	0.002	0.069	0.002
		SSC, D_b, OM	0.065	<0.001	0.065	0.001	0.066	0.001	0.067	0.002	0.068	0.002
	Neural network	SSC	0.067	<0.001	0.067	0.001	0.067	0.001	0.067	0.001	0.068	0.002
		SSC, D_b	0.065	0.001	0.066	0.001	0.065	0.001	0.067	0.002	0.067	0.003
		SSC, OM	0.063	0.001	0.064	0.001	0.065	0.002	0.065	0.002	0.067	0.004
		SSC, D_b, OM	0.063	0.001	0.063	0.001	0.064	0.001	0.065	0.002	0.066	0.003
Water retention at −15000 kPA	Nearest neighbor	SSC	0.051	<0.001	0.051	0.001	0.052	0.001	0.052	0.001	0.053	0.002
		SSC, D_b	0.051	<0.001	0.051	0.001	0.052	0.001	0.053	0.001	0.054	0.002
		SSC, OM	0.051	<0.001	0.051	0.001	0.052	0.001	0.052	0.001	0.054	0.002
		SSC, D_b, OM	0.051	<0.001	0.051	0.001	0.052	0.001	0.052	0.001	0.054	0.002
	Neural network	SSC	0.050	<0.001	0.049	0.001	0.049	0.001	0.049	0.001	0.050	0.001
		SSC, D_b	0.050	<0.001	0.050	0.001	0.050	0.001	0.050	0.001	0.050	0.002
		SSC, OM	0.048	<0.001	0.048	0.001	0.048	0.001	0.048	0.002	0.048	0.003
		SSC, D_b, OM	0.049	<0.001	0.048	0.001	0.048	0.001	0.049	0.002	0.049	0.003

† SSC = sand, silt, and clay content; D_b = bulk density; OM = organic matter content.

weight when calculating the final output. Differences in the estimations for such special cases may be the reason why the weighting method of Nemes et al. (2006) is, by a narrow and insignificant margin, still the most accurate method of the three.

Sensitivity of the Estimation Accuracy to the Number of Model Ensembles

We plotted the running RMSR values against the total number of ensembles after each replication data set had been applied to make estimations. We show a representative example of such plots in Fig. 3, for the SSCBDOM model and θ1500 as output. In Fig. 3, the average RMSR obtained after using M ensembles is shown, meaning that the last record on each line (at $M = 200$) in Fig. 3b matches the mean values shown in Table 2, line 12. For any M and $M + 1$ ensemble number pairs, differences among their performance were too small for the t-test to show statistically significant differences at $p = 0.95$. When M is small, e.g., $M < 20$, changes in average RMSR are visible in Fig. 3; however, due to large standard deviations, changes are not statistically significant.

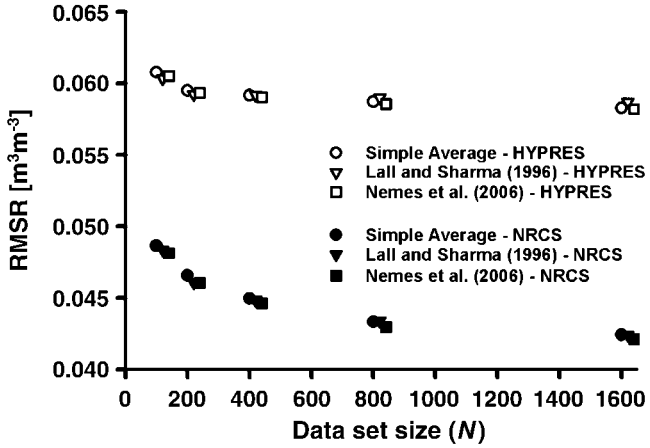

Fig. 2. Estimation accuracy using three different weighting methods to make estimations for the NRCS (below) and HYPRES (above) test data sets.

It was still desirable, however, to establish a minimum number of ensembles to be used to obtain stable RMSR, i.e., quasi-flat lines in Fig. 3. The case shown in Fig. 3

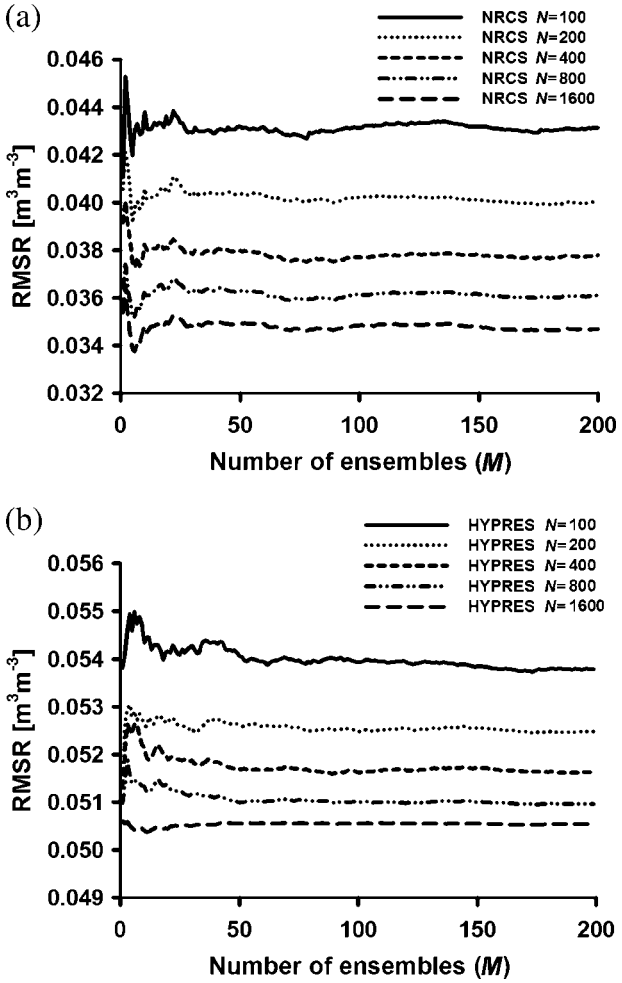

Fig. 3. Running root mean squared residuals (RMSR) for the (a) NRCS and (b) HYPRES test data sets for up to 200 ensembles using sand, silt, clay, bulk density, and organic matter content as input and water retention at −1500 kPa matric potential as output.

shows, for both test data sets, that the amplitude of changes is relatively larger using the first 30 (NRCS) to 50 (HYPRES) ensembles, and that there is practically no change after those cutoff points in any of the curves. For larger reference data set sizes, this flat tail of the graphs emerges after a smaller number of ensembles. Such behavior could be expected, since with larger subdata sets picked from the same master data set, the overlap between subdata sets will be more expressed, meaning that a smaller difference in estimations is expected, leading to less amplitude caused by an individual ensemble member. We have applied various criteria to determine the cutoff point at which we consider the estimations unchanged by the addition of further ensemble members. Table 3 lists cutoff ensemble numbers that were found after setting an absolute and a relative criterion to consider the estimations unchanged. In relative terms, we were seeking the ensemble number after which the change in RMSR did not exceed 0.01% of its value. Depending on the particular test data set and input–output combination used, it took the use of 21 to 55 ensembles to reach stability in this respect. In absolute terms, we were searching for the ensemble number from which the RMSR would remain within 0.001 $m^3 m^{-3}$ of the RMSR obtained using 200 ensemble members. This was achieved by using as few as two to seven ensemble members in many cases, and the maximum number of ensembles to meet this criterion was 26. Each value in Table 3 is the largest out of five values, representing the five different reference data set sizes. Typically, the presented values come from the models that used 100 or 200 samples in the reference data set because less overlap between data sets leads to larger variability in the estimations. Overall, it seems to be safe to state that, for this application, approximately 50 ensemble members are enough to minimize estimation errors.

Sensitivity of the *k*-Nearest Neighbor Technique to Data Density

We hypothesized that larger distances calculated according to Eq. [1] would lead to larger estimation errors. We correlated the absolute value of the estimation errors with the calculated distances of the nearest soils neighbors found in the reference data set to each individual target object. We used the absolute value of estimation errors in the analysis to make sure positive and negative errors did not cancel each other. Figure 4 shows the summary of our findings, where R^2 values have been averaged across reference data set sizes and input attribute sets used.

Correlation between the absolute value of errors and the distance of the kth neighbors is generally weak; the maximum value of R^2 was just above 0.1 when θ1500 was estimated for the NRCS test data set. Such correlations were consistently smaller for the HYPRES data set, regardless of which output attribute was estimated (maximum $R^2 = 0.056$). Another general observation is that it was the first nearest neighbor's distance that matters least for the magnitude of the estimation errors, regardless of which data set or output was used. This is suggested by the smallest R^2 found for $k = 1$ on each curve in Fig. 4. The importance of the distance of the other selected soils ($k = 2 \sim 25$) did not vary much, except for θ1500 of the NRCS set. There are substantial differences between R^2 values for the different kth neighbors, and the clustering of the points is noticeable.

The reason for such clustering is demonstrated in Fig. 5. As was mentioned above, points obtained for Fig. 4 are a result of averaging the outcomes of using different reference data set sizes. Figure 5 shows the results separately for the five reference data set sizes that yielded the "NRCS 1500" points in Fig. 4. As we followed the recommendation of Nemes et al. (2006) to optimize the number of selected soils (k), k is a function of the size of the reference data set that was used. Using the suggested equation—$k = 0.655N^{0.493}$, as above—we obtained values of 6, 9, 13, 18, and 25 for k for reference data set sizes of 100, 200, 400, 800, and 1600, respectively, which is reflected in Fig. 5. For this test data set (NRCS) and output variable (θ1500), we found larger variation in R^2 values obtained using different reference data set sizes than for other cases in Fig. 4. The R^2 values are more evenly distributed among neighbors for larger reference

Table 3. Number of ensembles needed for each input–output combination to reach a relative stability of <0.01% of the actual root mean squared residuals value if one more ensemble is added (Columns a) or an absolute stability of <0.001 $m^3 m^{-3}$ compared with the accuracy obtained using 200 ensembles (Columns b) for two different data sets. Each entry is the worst performing of five input data set sizes.

	NRCS				HYPRES			
	θ33‡		θ1500‡		θ33		θ1500	
Input attributes†	a	b	a	b	a	b	a	b
SSC	36	18	54	14	55	21	25	4
SSC, D_b	52	7	36	7	52	17	29	6
SSC, OM	21	2	30	14	39	26	24	6
SSC, D_b,OM	30	16	36	7	39	18	27	6

† SSC = sand, silt, and clay content; D_b = bulk density; OM = organic matter content.
‡ θ33 = water retention at −33 kPa matric potential; θ1500 = water retention at −1500 kPa matric potential.

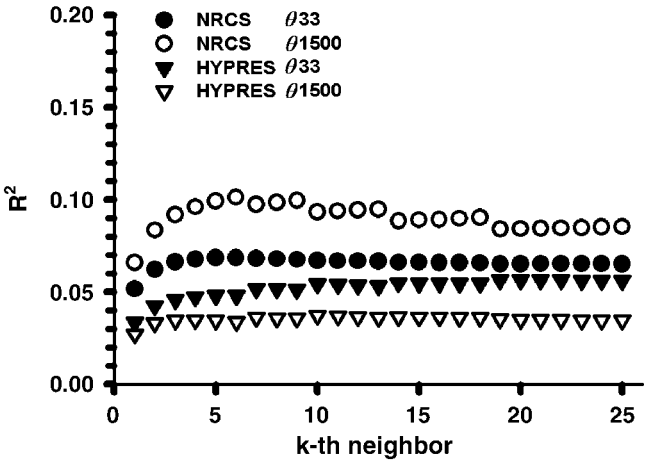

Fig. 4. Correlation between the absolute values of the estimation errors and the actual distance values (d_i) of each of the k neighbors. Different data set sizes are averaged.

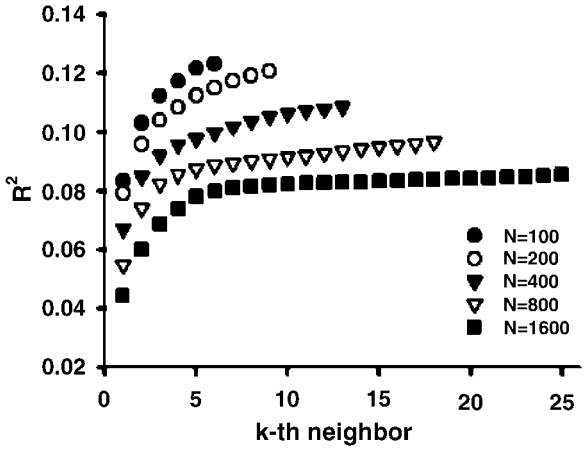

Fig. 5. Correlation between the absolute values of the estimation errors and the actual distance values (d_i) of each of the k neighbors. Different data set sizes are expanded for the NRCS data set for estimating water retention at −1500 kPa matric potential.

data set sizes, where more soils were selected. The largest correlation was found for the last (sixth) selected neighbor of the smallest reference data set (N = 100), R^2 = 0.125. Figure 6 shows the scatter plot and the regressed line that yielded this value in Fig. 5; all other data set input–output combinations yielded weaker correlations. Overall, the k-NN technique is not very sensitive to data density in terms of the estimation errors. Distant, but probably evenly distributed (surrounding), neighbors do not result in a much biased estimate for the individual soil. One selected neighbor may have a substantially larger value for the output attribute, but the other may have a smaller value by the same margin. The two values, when averaged, will yield a reasonable estimate and estimation error for the target soil, despite the large actual distances calculated for the neighbors, and the potentially largely different individual values in terms of the output attribute.

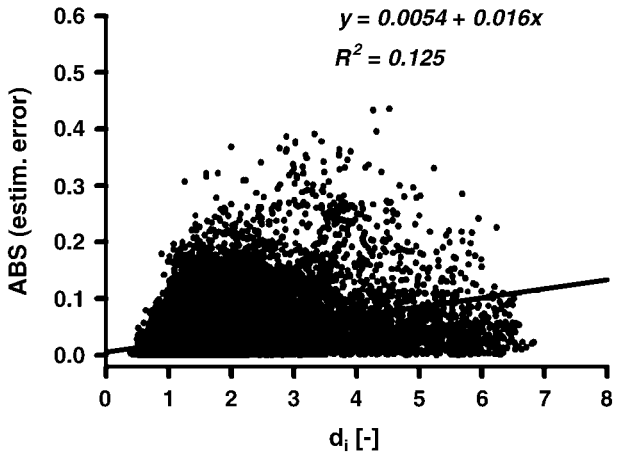

Fig. 6. Correlation between the absolute values (ABS) of the estimation errors and the distance values (d_i) of the sixth neighbor. Estimations are shown for estimating water retention at −1500 kPa matric potential for the NRCS test data set using 100 samples in the reference data set and all described inputs (sand, silt, clay content, bulk density, and organic matter content).

Sensitivity of the *k*-Nearest Neighbor Technique to the Potential Presence of Outliers

Figure 7 shows the sand and OM content of the soils used in this case study. The two outliers were added or omitted in each case according to the four different outlier settings. Figure 8 shows the RMSR values that were obtained using different input–output combinations. Outlier Setting 4 can be considered the point of reference in each case, as there were no outliers in any of the data sets in that setting. For θ33, there is practically no difference in the estimation results between any of the other outlier settings when cases using the same set of input attributes are compared. For θ1500, there is a slight ($<$0.002 m^3 m^{-3}) increase in the estimation errors with the outliers always being in the reference data set or being allowed to randomly be included in any of the two data sets. This difference is insignificant, however, compared with the variation within the ensemble RMSRs. In this case study, results in Outlier Setting 3—i.e., when the outliers are in the test data set—were not expected to differ much from the reference case. This is because the outliers represent only two out of 250 soils, and are thus practically a negligible part of the RMSR calculations. When they are present in the reference data set, however—in all cases or only randomly (i.e., Settings 1 and 2)—they have an impact on all calculations for all test soils because they have an impact on the ratio between different input attributes in the data standardization (see Eq. [2] and [3]). Our study, however, only tested the potential presence of outliers whose properties differed from the properties of the reference data set to a reasonable and realistic extent.

Unequal Input Attribute Weighting

Figure 9 shows the best combination of weights to minimize RMSR for the NRCS test data set. For both output variables, we show the running average weights to each input attribute, obtained considering the best 1 to 50 weight combinations. We examine the running averages of multiple models rather than individual mod-

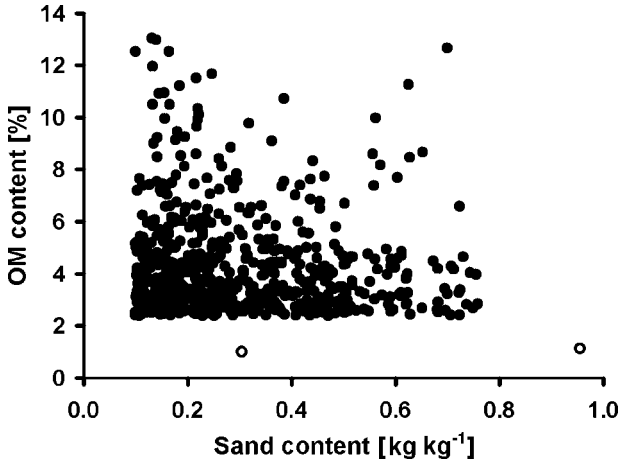

Fig. 7. Sand and organic matter (OM) contents in the limited NRCS data subset used in Case Study 5. Data points symbolized by open circles (N = 2) are considered hypothetical outliers.

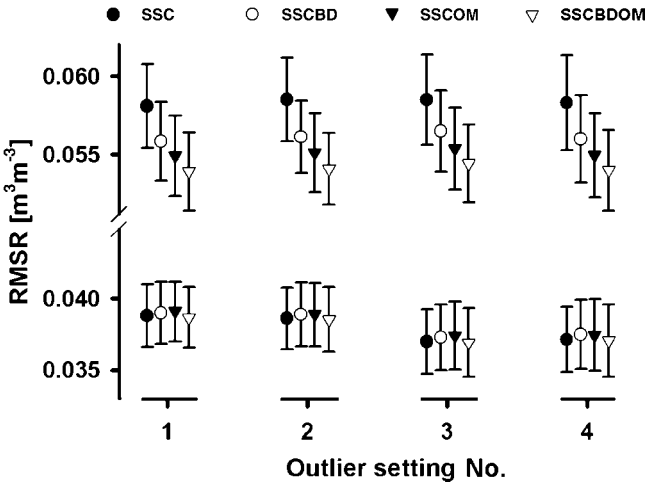

Fig. 8. Mean root mean squared residual (RMSR) values and their standard deviations for estimating water retention at (above) −33 kPa and (below) −1500 kPa matric potential obtained using different scenarios to include outliers in the data. Outlier settings: 1 = outliers always in the reference data; 2 = outliers mixed in randomly; 3 = outliers always in the test data; 4 = outliers not in the data. Vertical bars represent ±1 standard deviation, based on 200 ensembles. SSC = sand, silt, and clay content; SSCBD = sand, silt, clay, and bulk density; SSCOM = sand, silt, clay, and organic matter contents; SSCBDOM = sand, silt, clay, and organic matter contents and bulk density.

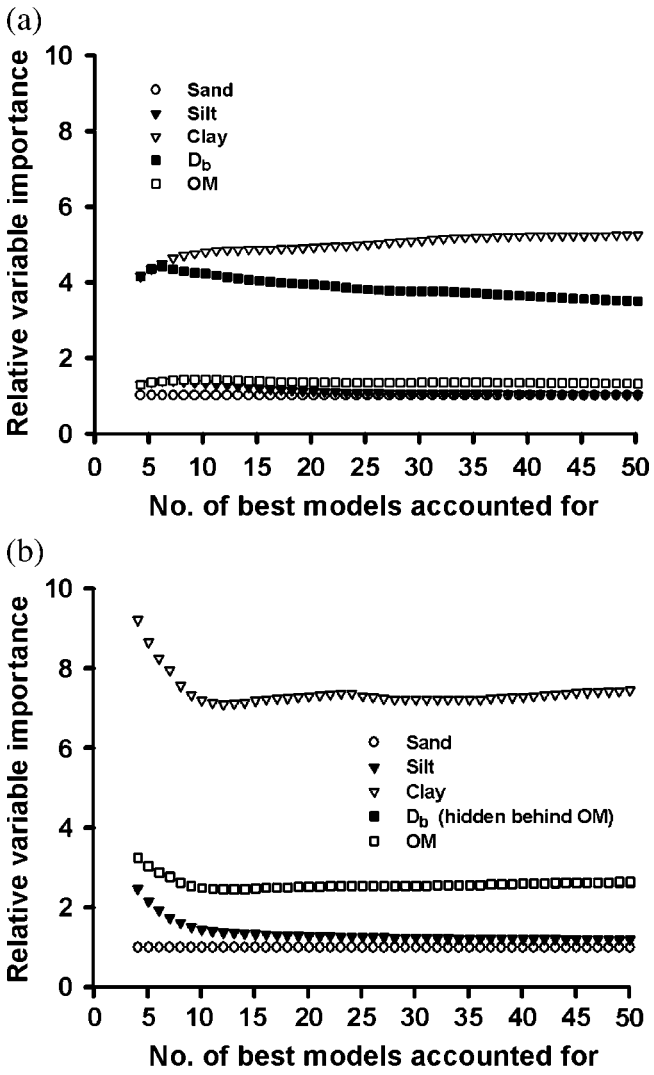

Fig. 9. Running input attribute weights to minimize estimation root mean squared residuals (RMSR) of (a) water retention at −33 kPa and (b) −1500 kPa matric potential. The model used sand, silt, and clay content, bulk density (D_b), and organic matter (OM) content as input and 1600 samples in the reference data set.

els to counter possible outlying cases and to monitor trends in the weighting of input attributes. It may also happen that the best combination of weights falls somewhere between our picks of scaling factors. In Fig. 9, weights are presented in a standardized way, so that the minimum weight assigned to any attribute will be equal to 1. That attribute will carry the least weight in the best performing models. In Fig. 9a, in the estimation of θ33, sand content was the least important input attribute, having its value at 1, regardless of how many of the first 50 best models we averaged. The weight of silt and OM content were also very close to 1. Clay content and D_b were two input attributes that had significantly larger weights assigned in the best models. They both started around a weight factor of 4, with D_b receiving gradually less weight but clay content assuming gradually more average weight when more of the best ranked models were considered. The weights of 3 to 5 that these two attributes assumed significantly differ from the weight of 1. In the estimation of θ1500, sand content was again the least important input attribute, having its value fixed at 1 throughout the examined spectrum of models. Silt content started with a weight close to 3, which quickly decreased to close to 1, as in the case for θ33. Bulk density and OM content carried practically the same weight throughout the examined spectrum of models. The weight of these attributes decreased quickly from 3.5 to a steady 2.7 to 2.8. Such quick decrease could also be observed in the case of clay content, the weight of which became steady at around 7.

In the estimation of both properties, the superior role of clay content became evident. As expected, it was more emphasized in the estimation of θ1500 than in the estimation of θ33. Bulk density is known to play an important role in determining the water-holding capacity of the soil closer to the wet end of the water retention curve, where soil structure has more influence through the formation of macro- and mesopores. Our findings correspond to this, with D_b having a larger impact on the estimation of θ33 than on the estimation of θ1500. Silt and sand content played a less important role; however, these properties are strongly correlated to clay content, the input attribute with the most weight.

Table 4 summarizes the results of estimations using unequal weighting of input attributes quantitatively. Once the best performing combination of attribute weights had been found for the NRCS test data set, we applied that combination to make estimations for the HYPRES test data set as well. The gain in estimation accuracy using unequal attribute weighting was marginal compared with the originally used equal weighting case, regardless of which test data set was used and which output attribute was estimated. The improvement in

Table 4. Comparison of root mean squared residuals (RMSR) obtained by the best combination of unequal input attribute weights and by equal input attribute weights for two different data sets.

	NRCS		HYPRES	
	θ33†	θ1500†	θ33	θ1500
RMSR minimum	0.0479	0.0329	0.0641	0.0497
RMSR with 1:1 weights	0.0491	0.0338	0.0649	0.0504

† θ33 = water retention at −33 kPa matric potential; θ1500 = water retention at −1500 kPa matric potential.

model performance always remained <0.0012 m^3 m^{-3} and did not yield a statistically significant improvement. This is an indicator of the relative insensitivity of this technique to suboptimal attribute weighting. Nemes et al. (2006) noted similar insensitivity when they examined other perspectives of finding the optimal settings for the *k*-NN model. Finding the technique insensitive to unequal weighting to a large extent also helps explain our findings in Case Study 5. Including outliers in the reference data set in practical terms means that we slightly change the weight of particular input attribute(s) compared with other input attributes. Finding relative insensitivity to unequal weighting points in the same direction as our findings in Case Study 5 about the simulated presence of outliers.

Estimations using Locally Specific Data

Results for all four test data sets are shown in Table 5 in terms of RMSR and MR for estimating θ33. Results for the estimation of θ1500 were similar and thus are not shown. Underlined data are data that were published earlier by Nemes et al. (2006) or in this study (c.f., Table 2). Estimation accuracy for the NRCS test data set did not change significantly with the addition of the BRAZ[OUT] data to the reference data set. Regardless of the number of input attributes used, changes to RMSR and MR remained well within the variation among ensemble members. Trends in RMSR and MR for the HYPRES data set were similar, except that the absolute values were somewhat larger than for the NRCS data set. This was, however, expected based on the findings of Case Study 1. Estimation accuracy for the HYPRES data set was not affected significantly by the addition of BRAZ[OUT] soils to the reference data set. A very small improvement (0.002–0.003 m^3 m^{-3}) was found when estimations were made for samples in the BRAZ[IN] data subset. The values in absolute terms are comparable—only slightly worse—to the values for the HYPRES test data set. This suggests that the *k*-NN technique is applicable for data sets that originate from other geographical areas, but whose data ranges are represented in the reference data set. The difference between the accuracy for the HYPRES and BRAZ[IN] data sets is expectedly due to the major difference in climatic conditions under which the BRAZ soils and soils of the NRCS and HYPRES data sets developed. For the BRAZ[OUT] test data set, estimations were significantly worse than for any other examined test data set, using only the NRCS data set as reference data. This was the case in terms of RMSR as well as MR. An MR of 0.1 m^3 m^{-3} is considered very large. With the inclusion of some BRAZ[OUT] data in the reference data set, estimation results changed dramatically for this test data set, as estimation accuracy in terms of RMSR became almost as good as for the NRCS data set (not significantly different at $p = 0.95$), and in terms of MR the improvement is 75 to 80% compared with the original values, depending on how many input attributes were used.

The hypothesis that using local data will improve estimations was not rejected by our experiment with these data. We showed that, while large improvements can be achieved for "locally specific" data, estimations for test data originating from other sources could remain practically unchanged. The fact that changes in the estimations were very limited for the BRAZ[IN] data set confirms that the change is not data origin specific, but specific for the actual soil properties. Test data set samples that fall to the edge of properties of nonspecific and locally specific soils or to the edge of the entire data domain still denote special cases. Presumably, many of the samples that fall to the edge of the data domain are the ones that show the most bias in the estimations. It is because these test samples are not uniformly surrounded by neighbors on each side within the data domain.

When parametric PTFs are used and (additional) locally specific data becomes available, a user either needs to develop his or her own independent PTF—for which there may not be enough data—or needs to add the data

Table 5. Summary of results, in terms of root-mean-squared residuals and mean residuals, of the estimation of water retention at −33 kPa matric potential for the NRCS, HYPRES, and BRAZ data sets. Underlined are numbers of comparison that have been published by Nemes et al. (2006) (for the NRCS set) or are included elsewhere in this study (for the HYPRES set, c.f., Table 2).

		NRCS		BRAZ(OUT)		BRAZ(IN)		HYPRES	
Reference data set	Input attributes†	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				RMSR, m^3 m^{-3}					
NRCS	SSC	0.054	0.003	0.113	0.005	0.072	0.001	0.069	<0.001
NRCS + BRAZ(OUT)	SSC	0.054	0.003	0.055	0.003	0.070	0.001	0.069	<0.001
NRCS	SSC, D_b, OM	0.050	0.002	0.112	0.004	0.072	0.001	0.065	<0.001
NRCS + BRAZ(OUT)	SSC, D_b, OM	0.051	0.002	0.053	0.003	0.069	0.001	0.065	<0.001
				MR, m^3 m^{-3}					
NRCS	SSC	0.000	0.003	−0.100	0.005	−0.012	0.001	0.004	0.001
NRCS + BRAZ(OUT)	SSC	0.001	0.003	−0.020	0.005	−0.009	0.001	0.004	0.001
NRCS	SSC, D_b, OM	0.002	0.002	−0.097	0.004	−0.009	0.002	0.008	0.001
NRCS + BRAZ(OUT)	SSC, D_b, OM	0.003	0.002	−0.025	0.005	−0.005	0.002	0.009	0.001

† SSC = sand, silt, and clay content; D_b = bulk density; OM = organic matter content.

and redevelop or readjust the original PTF. In the latter case, however, addition of data with differing data characteristics will change the final form of the relationships between inputs and the output. Such relationships—i.e., the equations of parametric PTFs—are valid globally for the entire range of soils on which they were developed. This way, estimations made for the entire range of soils are affected by the addition of some specific soils. This is not the case when the *k*-NN technique is used, given its design to use samples of the reference data set only from the neighborhood of each target object.

CONCLUSIONS

Nemes et al. (2006) introduced a *k*-NN technique to make estimations of water contents at −33 and −1500 kPa matric potentials. They found the performance of the technique to be comparable to that of NNets. In this study, we further tested the performance of the *k*-NN technique under different settings. Estimations were made for different data sets; the technique was exposed to different weighting methods, and its performance was monitored for different specific parts of the data space. We also examined the performance of this technique using different numbers of ensembles.

We found that the *k*-NN technique is, in general, insensitive to potentially suboptimal settings in many aspects. The *k*-NN technique performed almost equally well as NNet models developed on the same data to make estimations for data sets of different origin. The use of approximately 50 ensemble members resulted in estimation results that were not significantly affected by the addition of new ensemble members. The *k*-NN technique showed little sensitivity to the choice of applied sample weighting methods and to potential suboptimal settings in terms of input attribute weighting. Differences in data density in parts of the reference data set did not seem to substantially impact the estimation error. Substantial improvement was achieved for locally specific data when some local samples were included in the reference data set, while estimations for other samples remained almost unaffected.

The *k*-NN technique appears to be a competitive alternative to other techniques to develop PTFs. The technique shows a large degree of stability and insensitivity to nonoptimal algorithm settings and the use of different options. Differences introduced by such suboptimal settings and options in the algorithm or in data weighting caused only marginal changes in estimation accuracy. Such small differences are unlikely to have noticeable impact on simulation results that use one water retention estimate or the other. This technique can easily adopt new data without the need to redevelop equations, and can be developed into an "umbrella PTF" tool to make geographically or climatically specific estimations of soil hydraulic properties.

ACKNOWLEDGMENTS

We thank Dr. Javier Tomasella for making the Brazilian data collection available for this study.

REFERENCES

Baker, L. 2005. Optimisation of pedotransfer function models for soil hydraulic properties using an artificial neural network ensemble method. Ph.D. diss. Univ. of Abertay, Dundee, UK.

Dasarathy, B.V. (ed.). 1991. Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Comput. Soc. Press, Los Alamitos, CA.

Demuth, H., and M. Beale. 1992. Neural network toolbox manual. MathWorks, Natick, MA.

Efron, B., and R.J. Tibshirani. 1993. An introduction to the bootstrap. Monographs on statistics and applied probability. Chapman and Hall, New York.

Guber, A.K., Ya.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.

Harrold, T.I., A. Sharma and S.J. Sheather. 2003a. A nonparametric model for stochastic generation of daily rainfall occurrence. Water Resour. Res. 39:1300. doi:10.1029/2003WR002182.

Harrold, T.I., A. Sharma and S.J. Sheather. 2003b. A nonparametric model for stochastic generation of daily rainfall amounts. Water Resour. Res. 39:1343. doi:10.1029/2003WR002570.

Haykin, S. 1994. Neural networks, a comprehensive foundation. Macmillan College Publ. Co., New York.

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

Houtemaker, P.L., L. Lefaivre, J. Derome, H. Ritchie, and H.L. Mitchell. 1996. A system simulation approach to ensemble prediction. Monthly Weather Rev. 124:1225–1242.

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. Trans. ASAE 47:1437–1444.

Koekkoek, E.J.W., and H. Booltink. 1999. Neural network models to predict soil water retention. Eur. J. Soil Sci. 50:489–495.

Lall, U., and A. Sharma. 1996. A nearest-neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32:679–693.

MacLean, A.H., and T.U. Yager. 1972. Available water capacities of Zambian soils in relation to pressure plate measurements and particle size analysis. Soil Sci. 113:23–29.

Mehrotra, R., and A. Sharma. 2006. Conditional resampling of hydrologic time series using multiple predictor variables: A K-nearest neighbour approach. Adv. Water Resour. 29:987–999.

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

Minasny, B., A.B. McBratney, and K.L. Bristow. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93:225–253.

Molteni, F., R. Buizza, T.N. Palmer, and T. Petroliagis. 1996. The ECMWF ensemble prediction system: Methodology and validation. Q. J. R. Meteorol. Soc. 122:73–119.

Nemes, A., W.J. Rawls, and Ya.A. Pachepsky. 2006. Use of a non-parametric nearest-neighbor technique to estimate soil water retention. Soil Sci. Soc. Am. J. 70:327–336.

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

Nemes, A., J.H.M. Wösten, A. Lilly, and J.H. Oude Voshaar. 1999. Evaluation of different procedures to interpolate the cumulative particle-size distribution to achieve compatibility within a soil database. Geoderma 90:187–202.

Pachepsky, Ya.A., D. Timlin, and G. Várallyay. 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci. Soc. Am. J. 60:727–773.

Palmer, T.N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déqué, E. Díez, F.J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi et al. 2004. Development of a European multimodel ensemble system for seasonal to inter-annual prediction (DEMETER). Bull. Am. Meteorol. Soc. 85:853–872.

Schaap, M.G., and F.J. Leij. 1998. Database-related accuracy and uncertainty of pedotransfer functions. Soil Sci. 163:765–779.

Schaap, M.G., and F.J. Leij. 2000. Improved prediction of unsaturated hydraulic conductivity with the Mualem–van Genuchten model. Soil Sci. Soc. Am. J. 64:843–851.

Schaap, M.G., F.J. Leij, and M.Th. van Genuchten. 1998. Neural

network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J. 62:847–855.

Schaap, M.G., F.J. Leij, and M.Th. van Genuchten. 1999. A bootstrap–neural network approach to predict soil hydraulic parameters. p. 1237–1250. *In* M.Th. van Genuchten et al. (ed.) Proc. Int. Worksh., Characterization and Measurements of the Hydraulic Properties of Unsaturated Porous Media, Riverside, CA. 22–24 Oct. 1997. Univ. of California, Riverside.

Sharma, A., and R. O'Neill. 2002. A nonparametric approach for representing interannual dependence in monthly streamflow sequences. Water Resour. Res. 38:5. doi:10.1029/2001WR000953.

Soil Survey Staff. 1951. Soil survey manual. U.S. Dep. Agric. Handbook no. 18. U.S. Gov. Print. Office, Washington, DC.

Soil Survey Staff. 1997. National characterization data. Soil Survey Lab., Natl. Soil Survey Center, Lincoln, NE.

Tamari, S., J.H.M. Wösten, and J.C. Ruiz-Suárez. 1996. Testing an artificial neural network for predicting soil hydraulic conductivity. Soil Sci. Soc. Am. J. 60:771–774.

Tarboton, D.G., A. Sharma, and U. Lall. 1998. Disaggregation procedures for stochastic hydrology based on nonparametric density estimation. Water Resour. Res. 34:107–119.

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., Ya.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.

Wettschereck, D., D.W. Aha, and T. Mohri. 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif. Intell. Rev. 11(1–5):273–314.

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.

Yakowitz, S. 1993. Nearest-neighbor estimation for null-recurrent Markov time series. Stoch. Proc. Appl. 48:311–318.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek. 2003. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resour. Res. 39(7):1199. doi:10.1029/2002WR001769.

Ye, M., S.P. Neuman, and P.D. Meyer. 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour. Res. 40:W05113. doi:10.1029/2003WR002557.