Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters using Support Vector Machines

Navin K. C. Twarakavi*
Auburn Univ.
201 Funchess Hall
Auburn, AL 36849

Jirka Šimůnek
Dep. of Environmental Sciences
Univ. of California
Riverside, CA 92521

M. G. Schaap
Dep. of Soil, Water, and Environmental Science
Univ. of Arizona
Shantz Bldg.
Tucson, AZ 85721

Modeling flow in variably saturated porous media requires reliable estimates of the hydraulic parameters describing the soil water retention and hydraulic conductivity. These soil hydraulic properties can be measured using a wide variety of laboratory and field methods. Frequently, this proves to be an arduous task because of the high spatial and temporal variability of soil properties. In the last decade, researchers have shown a keen interest in developing a class of indirect approaches, called pedotransfer functions (PTFs), to overcome this problem. Pedotransfer functions predict soil hydraulic parameters using easily obtainable soil properties such as textural information, bulk density and/or few retention points. In this paper, we use a new methodology called Support Vector Machines (SVMs) to derive a new set of PTFs. Support vector machines represent a pattern recognition approach where the overall prediction error and complexity of the SVM structure are minimized simultaneously. We used the same database that was utilized to develop ROSETTA to generate the SVM-based PTFs. The performance of the SVM-based PTFs was analyzed using the coefficient of determination, root mean square error (RMSE) and mean error (ME). All soil hydraulic parameters estimated using the SVM-based PTFs showed improved confidence in the estimates when compared with the ROSETTA PTF program. Estimates of water contents and saturated hydraulic conductivities using the hydraulic parameters predicted by the SVM-based PTFs mostly improved compared with those obtained using the artificial neural network (ANN)-based ROSETTA. The RMSE for water contents decreased from 0.062 to 0.034 as more predictors were used, while the RMSE for the saturated hydraulic conductivity decreased from 0.716 to 0.552 (dimensionless $\log_{10}$ units). Similarly, the bias in the water contents estimated using the SVM-based PTF was reduced significantly compared with ROSETTA.

Abbreviations: ANN, Artificial neural networks; HYPRES, Database of hydraulic properties of European soils; ME, Mean Error; PTFs, Pedotransfer functions; RMSE, Root mean squared error; SLT, Statistical learning theory; SVMs, Support vector machines; UNSODA, Unsaturated soil hydraulic properties database; WISE, World inventory of soil emission potentials.

A number of hydrological models used to simulate situations ranging from field-scale water flow to global climate change rely on numerical techniques that simulate heat, water, and solute fluxes in the vadose zone. The use of flow models for variably saturated conditions requires accurate estimates of the hydraulic characteristics that govern water retention and water flow in soils (Wösten et al., 2001). The hydraulic characteristics of soils vary spatially from one location to another, and are also scale-dependent (Hopmans et al., 2002). The temporal variability can also occur as a result of various biological and human activities, such as root-growth, soil management and agricultural practices, or physical processes, such as soil shrinking-swelling, soil crusting, and/or water repellence (Wösten et al., 2001). Therefore, it is necessary to characterize the hydraulic characteristics of soils while keeping their spatial and temporal variability in perspective.

Ideally, it would be best to measure soil hydraulic parameters in the laboratory or field so that the variability in space and time can be sufficiently characterized. However, this difficult task is rarely accomplished because of the significant financial and time investments required for such measurements. Additionally, the spatial variability of soil hydraulic properties, their scale-dependency and possibly large modeling domains can make such characterization difficult.

To circumvent these practical difficulties with direct methods, researchers have shown keen interest in developing indirect methods beginning as early as 1912 (Briggs and Shantz, 1912). The basic idea behind these indirect methods, often called *pedotransfer functions* (PTFs after Bouma and van Lanen, 1987), is to predict *hard-to-measure* soil hydraulic properties (such as retention parameters and hydraulic conductivities) using *easily obtainable* input information (such as soil texture, bulk density, and particle-size distribution). In essence, PTFs represent predictive functions that translate the data *we have* (input) into the data *we need* (output) (Wösten et al., 2001)

Soil Sci. Soc. Am. J. 73:1443-1452
doi:10.2136/sssaj2008.0021
Received 23 Jan. 2008.
*Corresponding author (twarakavi@auburn.edu).

The complexity of existing PTF models varies from simple lookup tables that provide hydraulic parameters for particular textural classes (e.g., Carsel and Parrish, 1988; Wösten et al., 1995), to linear/nonlinear regression-based approaches (cf. Rawls and Brakensiek, 1985; Minasny et al., 1999), to models that incorporate physical relationships in soil water flow processes (e.g., Haverkamp and Parlange, 1986; Arya and Paris, 1981; Tyler and Wheatcraft, 1989; Nimmo et al., 2007). More sophisticated methods use various kinds of ANNs (Pachepsky et al., 1996; Tamari et al., 1996; Minasny et al., 1999), Group Method of Data Handling (Pachepsky and Rawls, 1999), or make use of multi-dimensional nearest neighbor techniques (Nemes et al., 2006a, 2006b). Wösten et al. (2001) provided a detailed review of the various PTFs that had been developed.

Pedotransfer functions are usually developed by examining the relationships between input data (textural properties) and soil hydraulic parameters (such as retention curve and/or hydraulic conductivity function parameters) from existing soil databases. Several large databases such as UNSODA (Leij et al., 1996), HYPRES (Wösten et al., 1999), and WISE (Batjes, 1996) are available for the development of PTFs. The accuracy and reliability of a given PTF approach is determined by how well the various relationships between the input data (i.e., information such as texture or bulk density) and the output data (i.e., soil hydraulic parameters) are represented in the PTF structure. These relationships often tend to be highly nonlinear, and are therefore not well established. Consequently, regression-based PTFs (cf. Rawls and Brakensiek, 1985; Minasny et al., 1999) have often performed poorly because they require adequate prior knowledge of these relationships.

Even though regression-based PTFs have been widely used in the past due to their simplicity, PTFs based on pattern recognition approaches (cf. Pachepsky et al., 1996; Tamari et al., 1996; Schaap et al., 2001) seem to have become more popular.

Pattern recognition approaches help describe the underlying relationships between the given inputs and outputs by 'learning' from a training data set. After training, pattern recognition methods lead to high-dimensional nonlinear functions. It can be said that pattern recognition approaches are mathematical models obtained in an experimental way. If there were no data (examples, patterns, and observations), there would be no learning, and consequently no pattern recognition tools. One such commonly used pattern recognition tool belongs to a class of methods called Artificial Neural Networks (ANNs).

An ANN is a classical pattern recognition paradigm inspired by the way biological nervous systems, such as the brain, process information (Hastie et al., 2001). Artificial Neural Networks are composed of a large number of highly interconnected processing elements (called neurons) working in unison to solve specific problems. The key element of ANNs is user-defined, that is, their complicated inter-woven structure. Because of this, an ANN's optimality for a specific problem is heavily influenced by user expertise, though this can be mitigated somewhat by exploring a number of different ANN topologies (e.g., Ye et al., 2007). Once the neural network structure is selected, the objective of learning from the training data in ANNs is to calculate the optimal weight for each of the links in the neural net by minimizing the overall prediction error. The optimal high-dimensional relation between the input data (such as particle size or bulk density) and output data (hydraulic parameters) is learned from a given set of training data (input-output response patterns).

While ANN-based PTFs have been relatively successful, there are a number of weaknesses that need to be considered in their development and application. Key weaknesses include: (i) ANNs have a number of coefficients (weights) that do not permit easy physical interpretation (Schaap et al., 2001), (ii) the ANNs's structure has to be selected a priori and therefore may not be optimal since there are many types of neurons and many types of possible connections (Wösten et al., 2001), (iii) a higher number of neurons and connections than required can result in overfitting and over parameterization (Hastie et al., 2001) and (iv) due to the complexity of the ANNs structure and the large number of weights that are being "trained" as the network "learns", there is no assurance that the learning algorithm will find optimum weights that minimize prediction errors. The procedure can get stuck in a local minimum, even though it can be overcome mostly. In lieu of the problems associated with ANN-based PTFs, there is need for a better pattern recognition tool to improve the PTFs accuracy and reliability. This has been duly recognized in a review by Wösten et al. (2001). Tamari et al. (1996) used radial basis functions to develop PTFs. Radial basis functions are an improvement over traditional ANNs but they suffer from many of the same weaknesses as ANNs.

Recently, a number of new pattern recognition tools have been proposed that aim to improve on the weaknesses of ANNs. For example, over-fitting can be mitigated by constraining the ANN optimization using independent data (Schaap, 2004) or by using bayesian regularization (MacKay, 1992; Ye et al., 2007). Support vector machines (SVMs) provide another promising approach which, unlike ANNs where the complexity of the structure is fixed a priori and only the prediction error can be minimized, represent a pattern recognition approach where the overall prediction error and complexity of the SVM structure are simultaneously minimized (Vapnik, 1995; 1998).

The SVM regression methodology (Vapnik, 1995, 1998) is based on the Statistical Learning Theory (SLT), which is a unique philosophy for addressing the problems and techniques of pattern recognition. The SLT proved that a pattern recognition method would generalize well (i.e., provide good performance on independent data) when the structural complexity of the pattern recognition method and the prediction error in the training sets were simultaneously minimized (Vapnik, 1995, 1998). Good performance on a training data set is a necessary but insufficient condition for a robust pattern recognition method.

Support vector machines have been used successfully for a wide array of classification and regression problems. Recently, a number of SVM applications have been introduced in hydrological sciences (Kanevski and Maignan, 2004; Tartakovsky and Wohlberg, 2004; Wohlberg et al., 2006). Dibike et al. (2001) applied SVM to remotely sensed image-processing problems and reported a superior performance over the traditional ANNs. Liong and Sivapragasam (2000) showed that SVM performance was superior to ANNs in forecasting flood stage. Wohlberg et al. (2006) showed that SVM methodology tends to improve the predictions in some case studies, especially in regression-based problems such as kriging. Asefa et al. (2004, 2005) successfully used the SVM methodology for a host of pattern recognition problems in hydrology, ranging from ground water modeling, to stream flow predictions. In this paper, we analyze the performance of the pattern recognition approach based on the SVMs for generating PTFs.

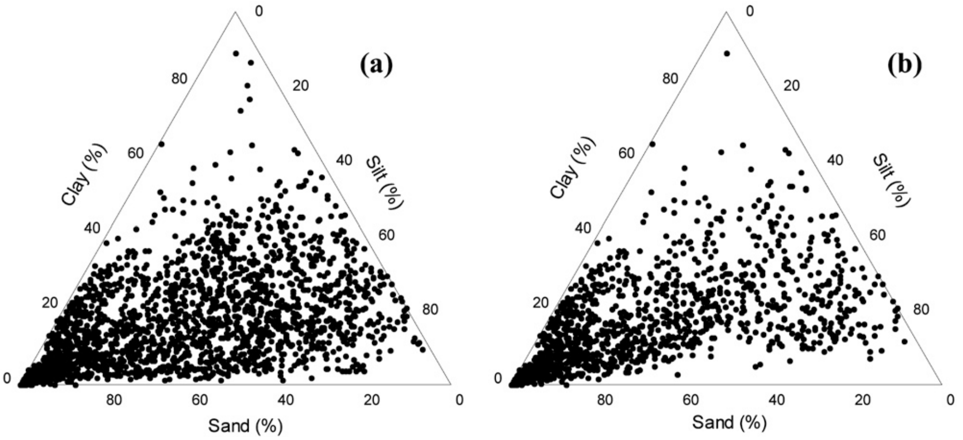

Fig. 1. Textural distribution of soil samples containing (a) retention data, and (b) saturated hydraulic conductivity.

MATERIALS AND METHODS

Soil Hydraulic Parameters

In this study, we consider the retention curve and hydraulic conductivity function described by the van Genuchten model (van Genuchten, 1980). Of all available models for representing soil hydraulic properties, the van Genuchten model is perhaps the most widely used model for characterizing hydraulic conductivity and water content dependence on the capillary pressure head. The van Genuchten model is described as follows.

$$\theta(h)=\begin{cases}\theta_r+\dfrac{\theta_s-\theta_r}{(1+|\alpha h|^n)^m} & h<0\\ \theta_s & h\geq 0\end{cases} \qquad [1a]$$

$$S_e(h)=\frac{\theta(h)-\theta_r}{\theta_s-\theta_r} \qquad [1b]$$

$$m=1-1/n, \qquad n>1 \qquad [1c]$$

where θ(h) is the volumetric water content, (L^3 L^{-3}) as a function of the pressure head h (L), θ_s and θ_r are the saturated and residual volumetric water contents, $S_e(h)$ is the effective soil water saturation (-) for the pressure head h (L), and α (L^{-1}) and n (-) are van Genuchten shape parameters. A combination of Eq. [1] with the Mualem pore-size distribution model (Mualem, 1976) yields the following expression for unsaturated hydraulic conductivity (van Genuchten, 1980):

$$K(S_e)=K_sS_e^L\left[1-(1-S_e^{\frac{1}{m}})^m\right]^2 \qquad [2]$$

where $K(S_e)$ is the unsaturated hydraulic conductivity (LT^{-1}) as a function of the soil water saturation (or pressure head), K_s is the saturated hydraulic conductivity (LT^{-1}), and L is an empirical parameter, often assumed to be 0.5 (Mualem, 1976). In this research, we develop PTFs for the following soil hydraulic parameters: (a) retention curve parameters (θ_s, θ_r, α and n) and (b) saturated hydraulic conductivity K_s.

Data Set

For developing the PTFs, we used the same data set used by Schaap and Leij (1998) and Schaap et al. (2001) for calibrating the ROSETTA model. Schaap and Leij (1998) assembled this set from a variety of available data sets. The assembled data set consists of data that are heterogeneous and diverse with respect to the measurement procedures, sources and scales of measurements. Figure 1 shows the textural distribution of the data set used in this study. Table 1 also lists the summary statistics for the retention parameters and the saturated hydraulic conductivity for different textural classes in the dataset. The dataset contained 2134 soil samples with water retention data. As each of these soil samples had at least six retention curve points, the entire data set included 20,574 θ(h) points. Measurements for saturated hydraulic conductivity (K_s) were available for 1306 of these soils samples.

The retention curve parameters were estimated for the 2134 soil samples with the retention data using a curve-fitting approach. The water

Table 1. Average water retention parameters, saturated hydraulic conductivities, and bulk densities for each textural class in the database. The values in parentheses represent the standard deviation (Schaap and Leij, 1998).

Class	Water retention						Saturated hydraulic conductivity		
	N^a	D_b†	θ_r	θ_s	log(α)	log(n)	N‡	D_b	log(K_s)
		g cm^{-3}	cm^3 cm^{-3}	cm^3 cm^{-3}	log(cm^{-1})			g cm^{-3}	log(cm d^{-1})
Sand	308	1.53(0.12)	0.053(0.029)	0.375(0.055)	−1.45(0.25)	0.50(0.18)	253	1.53(0.13)	2.81(0.59)
Loamy Sand	205	1.52(0.19)	0.049(0.042)	0.390(0.070)	−1.46(0.47)	0.24(0.16)	167	1.53(0.19)	2.02(0.64)
Loam	249	1.37(0.25)	0.061(0.073)	0.399(0.098)	−1.95(0.73)	0.17(0.13)	113	1.42(0.22)	1.08(0.92)
Sandy Loam	481	1.46(0.26)	0.039(0.054)	0.387(0.085)	−1.57(0.56)	0.16(0.11)	314	1.55(0.18)	1.58(0.66)
Silt Loam	332	1.28(0.27)	0.065(0.073)	0.439(0.093)	−2.30(0.57)	0.22(0.14)	135	1.42(0.14)	1.26(0.74)
Sandy Clay Loam	181	1.57(0.18)	0.063(0.078)	0.384(0.061)	−1.68(0.71)	0.12(0.12)	135	1.59(0.18)	1.12(0.85)
Silty Clay Loam	89	1.32(0.18)	0.090(0.082)	0.482(0.086)	−2.08(0.59)	0.18(0.13)	40	1.36(0.12)	1.05(0.76)
Clay Loam	150	1.42(0.19)	0.079(0.076)	0.442(0.079)	−1.80(0.69)	0.15(0.12)	62	1.44(0.23)	0.91(1.09)
Silt	6	1.33(0.09)	0.050(0.041)	0.489(0.078)	−2.18(0.30)	0.22(0.13)	3	1.39(0.03)	1.64(0.27)
Clay	92	1.39(0.20)	0.098(0.107)	0.459(0.079)	−1.82(0.68)	0.10(0.07)	60	1.40(0.23)	1.17(0.92)
Sandy Clay	12	1.59(0.10)	0.117(0.114)	0.385(0.046)	−1.48(0.57)	0.08(0.06)	10	1.60(0.08)	1.06(0.89)
Silty Clay	29	1.36(0.15)	0.111(0.119)	0.481(0.080)	−1.79(0.64)	0.12(0.10)	14	1.33(0.16)	0.98(0.57)

†Bulk density.

‡Number of soils per textural class.

retention data for each soil sample were fitted using an algorithm utilized by Schaap and Leij (1998) to estimate the following soil hydraulic parameters (θ_s, θ_r, α and n). Fitting was performed with the simplex or amoeba algorithm using the following constraints: $0.0 \le \theta_r \le 0.3$ cm^3 cm^{-3}, $0.6\phi \le \theta_s \le \phi$ cm^3 cm^{-3} (where ϕ is the total porosity), $0.0001 \le \alpha \le 1.000$ cm^{-1} and $1.001 \le n \le 10$. Measured saturated hydraulic conductivity was directly used as it was available for 1306 of the soil samples. The soil hydraulic parameters for each soil sample (that is, fitted retention curve parameters and the measured saturated hydraulic conductivity, where available) were then juxtaposed to other soil properties (such as textural information, bulk density, and water contents at particular pressure heads) to create the input data for the training and testing of PTFs. For improving the statistical robustness of the PTFs, logarithmically transformed values of the measured saturated hydraulic conductivities [*log*(K_s)] and the fitted retention curve shape parameters [log(α) and log(n)] were used (Schaap et al., 2001).

It is important to note that some textural classes (such as silt) do not have a sufficient number of samples in the data set, and thus calibrated PTFs for these textural classes are likely to be highly uncertain. It can be expected that the uneven distribution of soil samples in different textural classes will impact the reliability of the PTF predictions.

Theory of Support Vector Regression

The goal of a pattern recognition method is to estimate an unknown real-value variable using the mathematical functional relationship as follows:

$$y=\beta(\mathbf{x})+\delta \qquad [3]$$

where δ is an independent and identically distributed zero mean random error (noise), $\boldsymbol{x}$ is a multivariate input of m-dimensions, y is a scalar output and $\beta(\boldsymbol{x})$ is a function, likely nonlinear. The estimation of the $\beta(\boldsymbol{x})$ function is based on a finite number (n_t) of samples, the so-called training data set: $(\mathbf{x}_i, y_i)$ $(i = 1, ..., n_t)$.

The SVM regression is based on the generalized regression formulation. In the SVM regression, the input $\boldsymbol{x}$ is first mapped onto an m-dimensional space using some fixed (nonlinear) mapping, and then a linear regression model relating the input-output data is constructed. Mathematically, the SVM regression is formulated as:

$$f(\mathbf{x},\mathbf{w})=\sum_{j=1}^{m} w_j\, g_j(\mathbf{x})+b \qquad [4]$$

where $g_j(\boldsymbol{x})$, $j = 1..m$ denotes a set of nonlinear transformations, w_j are associated weights and b is the bias term.

Similar to traditional regression methods, an SVM regression model is developed by minimizing the error in predictions during the training stage. The quality of each prediction for a noisy data set is measured by a novel error function known as the ε-insensitive loss function (Vapnik, 1995; 1998):

$$L(y, f(\mathbf{x},w)) = \begin{cases} 0 & \text{if} \left|y - f(x,w)\right| \le \varepsilon \\ \left|y - f(x,w)\right| - \varepsilon & \text{otherwise} \end{cases} \qquad [5]$$

The value of ε in the loss function dictates the sensitivity of the transformation to noise in the data. A value of ε equal to zero essentially converts the ε-insensitive loss function into a least-modulus loss function and Eq. [5] into the mean absolute error. It has been shown in many studies that a non-zero ε results in a good SVM regression (Vapnik, 1995; 1998). Figure 2 shows a schematic of the ε-insensitive loss function. Once the residuals are estimated using the ε-insensitive loss function, the cumulative training error can be described as:

$$R(w) = \frac{1}{n_t}\sum_{i=1}^{n_t} L(y_i, f(x_i, w)) \qquad [6]$$

The SVM regression performs a linear regression using ε-insensitive loss function while trying to reduce the SVM's structural complexity by minimizing $\|w\|^2$ (Vapnik, 1995, 1998). This is achieved by introducing the non-negative slack variables ξ_i and ξ_i^*, $i = 1..n_t$ to measure the deviation of training samples outside the ε-insensitive zone (Fig. 2). Minimization of $\|w\|^2$ influences the structural complexity of the SVM regression model because it influences the final estimate of hyperparameters which govern the structure of SVM model. Thus, the SVM regression is formulated as a minimization of the following function:

$$\frac{1}{2}|\,w\,|^2 + C\sum_{i=1}^{n_t}(\xi_i + \xi_i^*) \qquad [7]$$

Subject to

$$\begin{cases} y_i - f(\mathbf{x}_i, w) - b \le \varepsilon + \xi_i^* \\ f(\mathbf{x}_i, w) + b - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, i = 1, ..., n_t \end{cases}$$

This optimization problem can be transformed into a dual problem (Vapnik, 1995), and its solution is given as:

$$f(\mathbf{x}) = \sum_{i=1}^{n_t}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b$$

such that

$$0 \le \alpha_i^* \le C, 0 \le \alpha_i \le C \qquad [8]$$

where n_t is the number of points in the training data set and $K(\mathbf{x}_i, \mathbf{x})$ is a kernel function.

A number of kernels are available (Vapnik, 1995, 1998). The kernel function is a symmetrical function that satisfies the so-called Mercer's conditions (Vapnik,

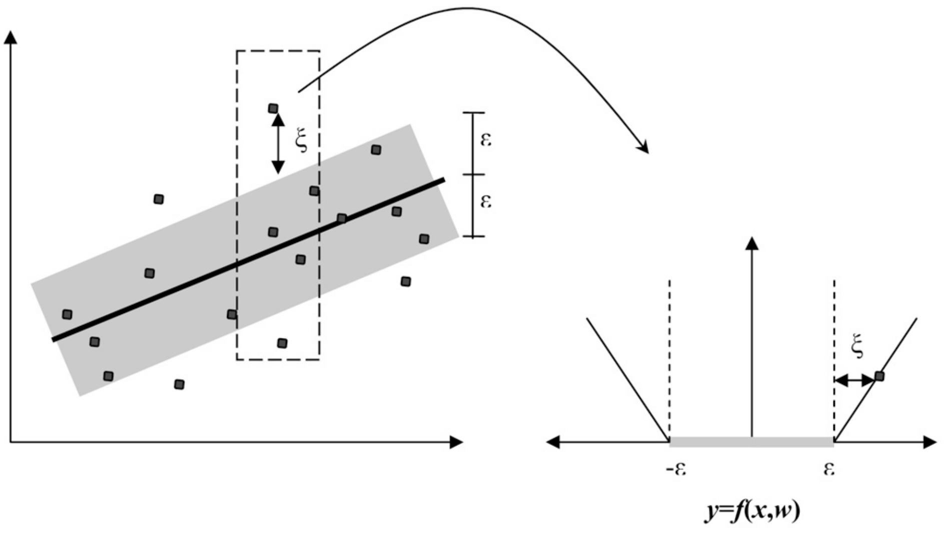

Fig. 2. Illustration of the SVM regression with the •-insensitive loss function. All the errors contributed by points within the •-tube are not considered during the SVM regression.

1995, 1998). The performance of SVMs depends at least as strongly on the choice of a kernel as on a kernel's parameterization (Wohlberg et al., 2006). In fact, the lack of a principled way to identify the optimal kernel is considered a main weakness of SVMs. In this study, we consider one of the most commonly used kernels, namely the Radial Basis kernel. The radial basis function is mathematically expressed in Eq. [9].

$$K(\mathbf{x}_i,\mathbf{x})=\exp\left(-\frac{\left\|x-x_i\right\|^2}{2\gamma^2}\right) \qquad [9]$$

where γ is the kernel parameter called as the kernel width. One may define the radial basis kernel by specifying the kernel width γ a priori. Readers are referred to Vapnik, (1995, 1998) for a detailed description.

During the training of SVM regression models, only certain data points in the training data set have non-zero coefficients in Eq. [8]. These points are also called as support vectors. The number of support vectors (n_{sv}) is a good indicator of how well the SVM model may perform on an unseen dataset. A larger percentage of support vectors will lead to over-fitting of the training data set and poor predictions in the testing data set, and a smaller percentage will lead to under-fitting. For optimally fitting SVM using a training data set, the percentage of support vectors should be around 50% (Hastie et al., 2001).

Support vector machine-based regression has been found to show an improved performance for some cases in the past (Wohlberg et al., 2006). The performance of the SVM regression depends on a good selection of the following so-called hyper parameters: cost (C), insensitivity value (ε) and the radial basis kernel width (γ). The cost parameter C determines the tradeoff between the complexity of the SVM structure and the prediction error of the training set. For example, if the cost C is set too large, the resulting SVM regression would give minimal importance to the necessity of minimizing the SVM structure complexity. This could result in over-fitting of the training data and poor generalization. The insensitivity parameter, ε, controls the width of the insensitive zone. Larger values of ε lead to smaller numbers of support vectors, which can result in poor generalization.

Figure 3 shows a schematic of a trained SVM structure, and the steps involved in estimating the final output value, $f(\boldsymbol{x_t},w)$, given an input vector $\boldsymbol{x_t}$. The input vector $\boldsymbol{x_t}$ and the support vectors $\boldsymbol{X_i}$ (i = 1,.., n_{sv}) are nonlinearly transformed using the pre-defined kernel function and then linearly combined using weights w_i and the bias b obtained during training to estimate the final output value.

In this research, the SVM regression was performed in two stages: (i) training and (ii) testing. Before training, the training data was standardized for zero mean and unit variance. The standardized training data was used to develop the SVM regression models. The mean and standard deviation values used for standardizing, was used in the later stages for de-standardizing the predictions. The training stage aims at finding the optimal estimates of cost, C, insensitivity value, ε and the radial basis kernel width, γ, to achieve the best generalization. During the testing stage, the ability of the trained SVM to predict final values is evaluated.

In the past, optimal estimations of the SVM hyper-parameters were performed using the following three procedures: (i) based on a priori knowledge and user expertise, (ii) using a thorough grid-based search approach, and (iii) using an analytical estimation based on the statistical properties of the training data. In this study, we have opted to use the grid-based search approach. The objective of the grid-search method is to obtain the optimal hyper-parameters by estimating the error in the predictions for a training data set for every possible combination of the hyper-parameters within a feasible hyper-parameter space. The hyper-parameter which results in a minimum training error was chosen as optimal. To estimate the training error, we used a five-fold cross-validation approach. Readers are referred to Hastie et al. (2001)

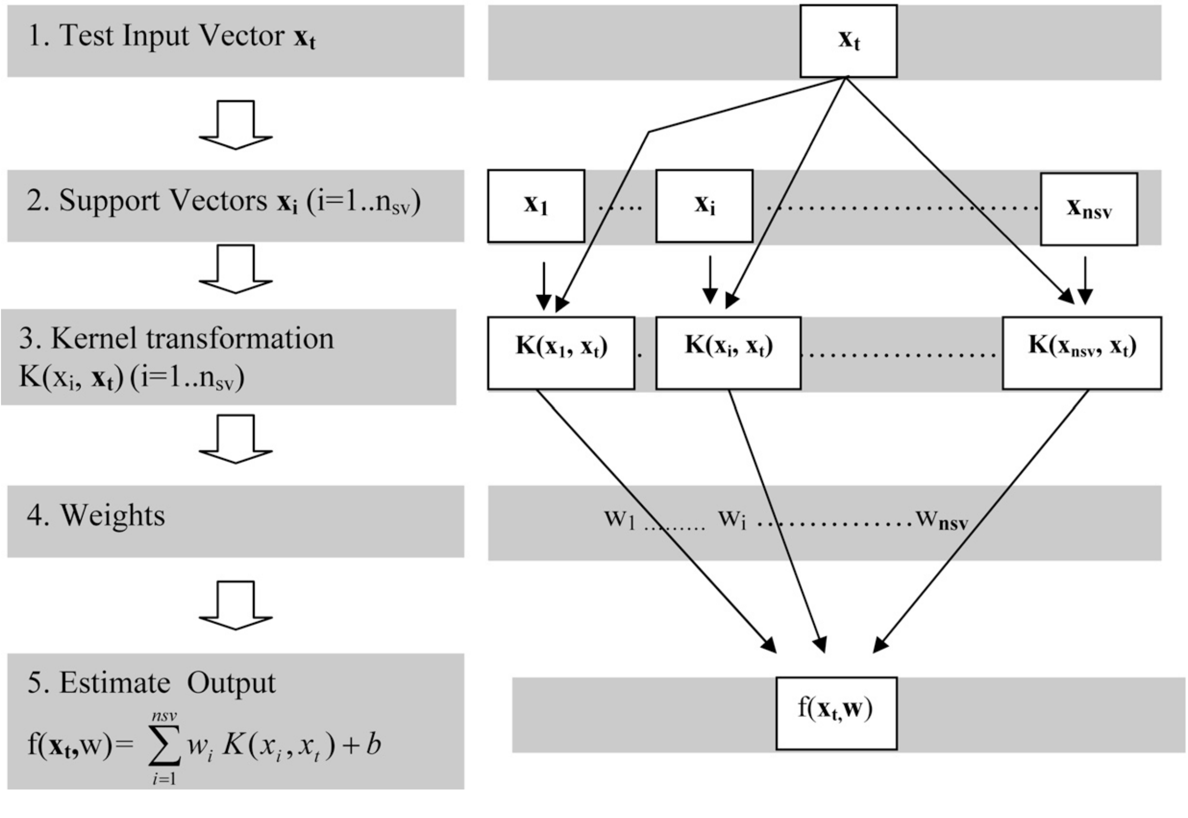
1. Test Input Vector $\mathbf{x_t}$
2. Support Vectors $\mathbf{x_i}$ (i=1..n_{SV})
3. Kernel transformation
K(x_i, $\mathbf{x_t}$) (i=1..n_{SV})
4. Weights
5. Estimate Output
$f(\mathbf{x_t},w)= \sum_{i=1}^{nsv} w_i K(x_i,x_t)+b$
$\mathbf{x_t}$
$\mathbf{x_1}$
$\mathbf{x_i}$
$\mathbf{x_{nsv}}$
K($\mathbf{x_1}$, $\mathbf{x_t}$)
K($\mathbf{x_i}$, $\mathbf{x_t}$)
K($\mathbf{x_{nsv}}$, $\mathbf{x_t}$)
w_1 .... w_i .... $w_{\mathbf{nsv}}$
f($\mathbf{x_t}$,$\mathbf{w}$)

Fig. 3. Flowchart showing the steps involved during estimation of output when input data is fed to a trained SVM regression model. Note that the Steps (2), (3) and (4) are performed at the same instance when the appropriate kernel such as the radial basis kernel is used.

for a detailed explanation of the five-fold cross-validation approach. The optimal hyper-parameters for the SVMs were estimated by searching within the feasible parameter space. The feasible parameter space for each of the hyper-parameters was constructed using a set of minimum and maximum possible values that were assumed to be the following a priori ($0.0001 < C < 1000$, $0.0001 < \gamma < 100$, $0 < \varepsilon < 1$). We assumed the aforementioned range of hyper-parameters to be appropriate based on number of previous studies (e.g., Asefa et al., 2004; 2005). A mesh increment of 0.01 was selected for the grid search to ensure optimality of hyper-parameters and the computational efficiency.

Performance Criteria

At different stages of the PTF development, it is required to quantify the amount by which an estimated value differs from the 'true' value of the quantity being estimated. Such quantification describes how well the estimator describes the 'true' values. In this research, such differences between estimated values and the 'true' values are quantified using the following performance criterion: (i) root mean square error (RMSE), (ii) mean error (ME), and (iii) coefficient of determination (R^2).

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\zeta_i-\zeta_i^{'})^2} \qquad [10]$$

$$\mathrm{ME}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\zeta_i-\zeta_i^{'})} \qquad [11]$$

$$R^2=1-\frac{\sum_{i=1}^{N}(\zeta_i-\zeta_i^{'})^2}{\sum_{i=1}^{N}(\zeta_i-\bar{\zeta}_i)^2} \qquad [12]$$

where N is the number of values, ζ_i $(i = 1..N)$ is the 'true' value of the quantity, $\bar{\zeta}$ is the mean of the 'true' values and ζ'_i $(i = 1..N)$ are the corresponding estimated values.

Support Vector Machine Methodology

We developed new PTFs based on the SVM methodology for the saturated hydraulic conductivity [$\log(K_s)$] and the four parameters of the water retention function [θ_s, θ_r, $\log(\alpha)$, and $\log(n)$]. The R statistical language was used for developing the SVM-based PTFs. Four SVM-based PTFs were developed for each set of soil hydraulic parameters by varying the input data predictors. The first set of PTFs (called S1) used sand, silt, and clay percentages as input predictors. The second set of PTFs (called S2) used bulk density, and sand, silt, and clay percentages as predictors, while the input data of the third set of PTFs (called S3) also included the water content at the pressure head of −330 cm. The last model (called S4) included the water content at the pressure head of −15,000 cm in addition to those predictors used for the third set of PTFs (S3). Pedotransfer functions for retention parameters were developed by training with the corresponding parameter data that were estimated by fitting the van Genuchten retention model to retention data of the soils (2134 soil samples). On the other hand, PTF for saturated hydraulic conductivity was trained using the measured saturated hydraulic conductivity values (1306 soil samples). Therefore, performance analysis of the PTFs using RMSE, R^2 and ME (Equations 10–12) was done by comparing the predicted values with fitted estimates [in case of θ_s, θ_r, $\log(\alpha)$, and $\log(n)$] and measured values [in case of $\log(K_s)$].

Pedotransfer functions were developed by combining the SVM approach with the bootstrap method (Efron and Tibshirani, 1993). In the bootstrap method, a single dataset of size 'N' is randomly resampled with replacement to create 'B', statistically similar subsets also of size 'N', while each of the 'B' subsets contain about 63% of the parent data set (Schaap et al., 2001). Using each of these 'B' subsets (also called as training data set), PTFs were first trained. The trained PTFs were then tested using the 37% of the parent dataset that was not represented in the subset. Incorporating the bootstrapping method into the generation of PTFs has several advantages: (i) it allows for estimating uncertainty in the PTF prediction, and (ii) it creates complementary testing and training data sets. A 'B' value of 60 was used in our calibration, following Efron and Tibshirani (1993) who suggested a B value between 50 and 100 for a robust model. Each of the SVM-based PTFs comprised of 60 SVMs that were calibrated using the bootstrapped subsets of the original dataset. Predictions of the soil hydraulic parameters from 60 SVMs were first averaged to report the mean parameter estimates and their standard deviations. The mean value of the soil parameters from the 60 bootstraps were reported as the predictions from the PTFs.

For each of the PTFs used to predict the water retention parameters, we also compared the predicted values in the training and testing data sets with the corresponding fitted values for all of the 60 bootstraps using the RMSE criterion. In case of the PTF developed to predict the saturated hydraulic conductivity, we compared predicted values in the training and testing data sets with the corresponding measured estimates for all of the 60 bootstraps using the RMSE criterion. The mean of the RMSE in the training and testing data sets of the 60 bootstraps for each PTF was reported as the training and testing error.

It is of interest to analyze how the errors in the predicted soil hydraulic parameters translated in the accuracy of the water contents that may be estimated using these predicted parameters (Eq. [1a–1c]). For this purpose, we calculated RMSE and ME values between water contents estimated using parameters predicted by PTFs (S1-S4) and the measured water contents [20,574 $\theta(h)$ points]. To understand how these RMSE and ME values varied for water contents ($\theta(h)$) as a function of pressure head (h), we also calculated these values for water contents grouped using 10 suction classes between 0, 3.2, 10, 32, 100, 320, 1000, 3200, 10 000, 32000, and 10000 cm. It is important to note that water contents estimated by using parameters predicted by PTFs cannot be more accurate than those estimated by using fitted parameters that were used to develop the PTFs. The RMSE and ME were also estimated between the water contents estimated using the fitted parameters and the measured water contents for comparison purposes.

RESULTS AND DISCUSSION

Development of Support Vector Machine-based Pedotransfer Functions

The four SVM-based PTFs (S1, S2, S3, and S4) were developed using the procedure described above. Table 2 lists the summary statistics for the hyper-parameters of all PTFs, along with the percentage of training data set that represents support vectors. One may view support vectors as those records in the training data set providing the maximum value to the SVM method. It may be seen from Table 2 that all developed PTFs have an optimal number of support vectors around 50% of the training dataset which is also the recommended value (Hastie et al., 2001). The optimal percentage of support vectors was also reflected in the performance of SVMs on the training and testing data sets.

Table 2. Summary statistics of estimates of hyperparameters and percentage of support vectors (SV) for the SV machines (SVM)-based Pedotransfer functions (PTFs) that predict water retention parameters and saturated hydraulic conductivity using different sets of input data. The estimated means and standard deviations (in parentheses) of all bootstraps in a PTF are reported. SSC– sand, silt, and clay percentages, D_b– bulk density, θ_{330} and θ_{15000}– water contents at pressure heads of −330 and −15,000 cm, respectively.

Parameter	PTF Model	Input data	Cost, C	Kernel width γ	Insensitivity ε	Number of SV
				(-)	(-)	%
θ_r	S1	SSC	9.2 (3.35)	0.28 (0.1)	0.48 (0.14)	57.55 (8.76)
θ_s			23 (2.1)	1.01 (0.05)	0.62 (0.2)	42.53 (16.05)
$\log_{10}(\alpha)$			9.4 (2.97)	0.45 (0.35)	0.68 (0.3)	38.74 (21.61)
$\log_{10}(n)$			22.6 (1.67)	0.27 (0.09)	0.5 (0.09)	41.61 (8.76)
$\log_{10}(K_s)$			16.6 (2.44)	0.82 (0.32)	0.41 (0.25)	57.17 (23.03)
θ_r	S2	SSS+D_b	6.2 (3.35)	0.4 (0.1)	0.7 (0.14)	41.3 (8.76)
θ_s			4.4 (2.1)	0.27 (0.05)	0.52 (0.2)	34.96 (16.05)
$\log_{10}(\alpha)$			4.2 (2.97)	0.67 (0.35)	0.74 (0.3)	33.02 (21.61)
$\log_{10}(n)$			13.6 (1.67)	0.26 (0.09)	0.5 (0.09)	40.35 (8.76)
$\log_{10}(K_s)$			16.6 (2.44)	0.16 (0.32)	0.51 (0.25)	43.83 (23.03)
θ_r	S3	SSC+ D_b + θ_{330}	16.6 (3.35)	0.16 (0.18)	0.51 (0.15)	43.83 (9.5)
θ_s			9.4 (2.79)	0.16 (0.3)	0.41 (0.11)	42.95 (8.5)
$\log_{10}(\alpha)$			20 (3.56)	0.16 (0.45)	0.4 (0.2)	48.86 (10.92)
$\log_{10}(n)$			8.8 (1.51)	0.28 (0.12)	0.4 (0.07)	43.31 (4.53)
$\log_{10}(K_s)$			9 (1.34)	0.28 (0.11)	0.43 (0.22)	43.2 (19.12)
θ_r	S4	SSC+ D_b + θ_{330} + θ_{15000}	13.3 (3.29)	0.17 (0.11)	0.51 (0.22)	43.53 (19.12)
θ_s			8 (4.83)	0.22 (0.13)	0.42 (0.08)	40.41 (9.55)
$\log_{10}(\alpha)$			9 (2.26)	0.34 (0.07)	0.29 (0.11)	56.86 (11.05)
$\log_{10}(n)$			20 (3.9)	0.27 (0.3)	0.19 (0.09)	58.41 (7.76)
$\log_{10}(K_s)$			6.4 (1.41)	0.21 (0.15)	0.37 (0.2)	45.13 (21.05)

As described in the methodology section, we estimated the RMSE of the predictions in the training and testing data sets for all the PTFs (Table 3). Table 3 shows that the performance of the PTFs for the training and testing data is very close, suggesting a near optimal training of the PTFs

To enable direct comparisons with Schaap et al. (2001), we used the developed PTFs to predict soil hydraulic parameters for the entire data set. Comparisons were made using RMSE and R^2. Table 4 lists the RMSE and R^2 values of the soil hydraulic predictions by the SVM-based PTFs for the entire dataset. One obvious observation is that R^2 values between fitted and PTF-predicted soil hydraulic parameters increase and RMSE decreases as more input predictors are used in the PTFs (from S1 to S4). This observation suggests that all predictors in the input data set have some information that is useful for estimating the hydraulic parameters. In all of the PTFs, correlations between predictions of θ_r and corresponding fitted estimates is the weakest among the predicted parameters. Bulk density seems to considerably improve predictions of θ_s (compare S2 and S1). On the other hand, the accuracy of predictions also seems to increase significantly (especially for θ_r, log (α) and log(n)) when retention curve data are used as additional predictors (compare S2 to S3 and S4).

Table 3. Root mean squared error (RMSE) of the water retention parameters and saturated hydraulic conductivities predicted by different support vector machine (SVM)-based pedotransfer functions (PTFs) for the training and testing data sets.

Parameter	Data	S1	S2	S3	S4
θ_r	Training	0.063	0.061	0.052	0.039
	Testing	0.072	0.07	0.064	0.045
θ_s	Training	0.075	0.052	0.049	0.047
	Testing	0.081	0.055	0.052	0.05
log(α)	Training	0.558	0.531	0.47	0.342
	Testing	0.567	0.537	0.476	0.348
log(n)	Training	0.132	0.127	0.11	0.08
	Testing	0.132	0.13	0.113	0.083
log(K_s)	Training	0.714	0.662	0.561	0.55
	Testing	0.72	0.662	0.57	0.556

We also compared the R^2 values for the predicted soil hydraulic parameters with the corresponding values for the ANN-based ROSETTA which is given in Schaap et al. (2001). It was noted that the RMSE and R^2 values suggest a significant improvement in predictions obtained by all the SVM-based PTFs when compared with the ANN-based ROSETTA. Compared with ROSETTA, SVM-based PTFs, using SSC as the input, shows the most improvement. Also, the extent of improvement in the soil hydraulic parameters predicted by SVM seems to be lesser and lesser with the number of inputs used in the PTFs.

Table 5 shows RMSE between the observed water contents and those estimated by using parameters predicted by ROSETTA and SVM-based PTFs. The RMSE estimates in Table 4 also indicate SVM-based PTFs are more accurate in predicting water contents. One may attribute this improvement in predictions due to one or both of the following reasons: (i) the improvement of SVM over ANNs or (ii) the superior structure of SVM-based

Table 4. Root mean squared error (RMSE) and R^2 estimates for the water retention parameters and saturated hydraulic conductivities predicted by different support vector machine (SVM)-based pedotransfer functions (PTFs) for the entire dataset.

Parameter	Performance measure	S1	S2	S3	S4
θ_r	R^2	0.282	0.38	0.48	0.79
	RMSE	0.066	0.064	0.056	0.041
θ_s	R^2	0.44	0.787	0.816	0.831
	RMSE	0.077	0.053	0.05	0.048
log(α)	R^2	0.497	0.567	0.684	0.848
	RMSE	0.561	0.533	0.472	0.344
log(n)	R^2	0.686	0.705	0.791	0.897
	RMSE	0.132	0.128	0.111	0.081
log(K_s)	R^2	0.681	0.737	0.817	0.826
	RMSE	0.716	0.662	0.564	0.552

PTFs over ROSETTA. The ANN-based ROSETTA predicts all four VG parameters using ONE single model (i.e., the ANN has four outputs corresponding to different parameters). In the case of the SVM-based PTFs developed here, there are four different models, each predicting one output (multi-model). It is our opinion that the multi-model approach provides SVM-based PTFs more flexibility, and could be one of the reasons why the SVM-based PTFs have lower prediction errors. To check if the observed improvement was due to multi-model nature of SVM-based PTFs or due to the superior structure of SVM compared with ANN, we need to do further research comparing single-model ANN, multi-model ANNs, single-model SVM and multi-model SVMs. This is a topic for future research.

Uncertainty

Estimations of uncertainty (such as standard deviation) in predictions by PTFs can be very useful, especially for further statistical and modeling studies. The user may view the uncertainty estimate as a measure of PTF's reliability (e.g., Schaap, 2004). Since we used bootstrapping for the SVM-based PTFs, we can easily estimate standard deviations for the predicted soil hydraulic parameters. Figure 4 shows the distribution of mean (μ) soil hydraulic parameters predicted by the S2 PTFs and the associated uncertainty (σ, standard deviation) as a function of soil texture for a specified bulk density of 1.3 g cm^{-3}. Here, we use the standard deviation to represent the uncertainty under the assumption that the errors are distributed normally, which may not necessarily be true. As already observed for ROSETTA by Schaap et al. (2001), the textural dependence of predicted parameter values concur with established knowledge. For example, θ_r increases for finer textures, while $\log(K_s)$ increases for coarser textures. Standard deviations for the predicted values of soil hydraulic parameters show similarity. They are especially high for textures that are less well represented in the input data set (compare Fig. 1 with Fig. 4). This was consistently observed for predictions by all PTFs (S1-S4), indicating that gathering more information for textures that are less represented in the training dataset (such as silt) is a necessity.

Water Contents and Saturated Hydraulic Conductivity

The soil hydraulic parameters predicted using different SVM-based PTFs were used to estimate water contents at different pressure heads with the van Genuchten model. These predicted water contents were then compared with the corresponding measured values using RMSE. Table 5 summarizes RMSE of water contents and saturated hydraulic conductivities predicted using ROSETTA and the SVM-based PTFs for different input predictors. For saturated hydraulic conductivities, RMSE of predictions by ROSETTA and SVM are comparable for lower-order PTFs which use only textural information and bulk density. However, when retention data points are also included into the PTFs training (as in S3 and S4), the SVM-based PTFs show considerable improvement. Compared with ROSETTA, RMSE for water contents predicted by the SVM-based PTFs decreased by 10 to 25% for different PTFs. Compared with other PTFs, Table 5 indicates that the hydraulic conductivities predictions by SVM-based PTFs show only a marginal improvement over ROSETTA.

Table 5. Root mean square error (RMSE) estimates between the measured and corresponding predicted water contents and saturated hydraulic conductivities by support vector machine (SVM)-based pedotransfer functions (PTFs) and ROSETTA.

PTF Model	Input data used†	PTF	$\theta(h)$	$\log(K_s)$
S1	SSC	SVM-based	0.062	0.716
		ROSETTA	0.076	0.717
S2	SSC+D_b	SVM-based	0.053	0.662
		ROSETTA	0.068	0.666
S3	SSC+ D_b + θ_{33}	SVM-based	0.038	0.564
		ROSETTA	0.047	0.586
S4	SSC+ D_b + θ_{33}+θ_{15000}	SVM-based	0.034	0.552
		ROSETTA	0.044	0.581
Direct fit			0.012	NA

†SSC, sand silt and clay; Db, bulk density.

Figure 5 shows RMSE between water contents estimated using parameters predicted by PTFs (S1-S4) and the measured water contents as a function of the pressure head. A similar analysis was performed by Schaap et al. (2001). The figure also shows the RMSE for the water contents estimated by using the fitted parameters. Naturally, one may expect that the water contents predicted by the PTFs cannot be more accurate than those estimated by using the fitted parameters. Compared with the results of ROSETTA (Schaap et al., 2001), the SVM-based PTFs show improvements for different pressure head intervals. Out of all the PTFs, only S4 performs better at lower pressure heads ($<-32{,}000$ cm), while all other PTFs seem to have poor performance at these pressure heads. It is important to note that we had only 28 points below the pressure head of $-32{,}000$ cm. Thus, the poor performance of PTFs at lower pressure heads could be a statistical problem.

Figure 6 depicts the ME for water contents estimated using the parameters predicted by the SVM-based PTFs as a function of pressure head. The ME for the water contents estimated using the RETC-fitted parameters is also shown. Clearly, the RETC-fitted parameters show minimal ME indicating minimal bias in the estimated water contents. The water contents estimated using the parameters predicted by the SVM-based PTFs show relatively more bias. The water contents are overestimated near saturation (especially when $\log(h) < 0.5$ cm) and underestimated at lesser saturations. It was observed that this trend is very similar to ME of the water contents estimated by using ROSETTA (Schaap et al., 2001). However, the absolute ME value of the water contents estimated using parameters predicted by SVM-based PTFs indicate lesser bias than those estimated using parameters estimated by ROSETTA. Similarly in Fig. 6, the ME of the estimated water contents seems to be reduced as more predictors are used in the corresponding PTF.

SUMMARY AND CONCLUSIONS

The SVM methodology was successfully applied to develop PTFs that used different input predictors to estimate soil hydraulic parameters. These PTFs utilize some or all of the following predictors: sand, silt, and clay percentages, bulk density and retention data at one or two points (at the field capacity and the wilting point). Bootstrapping was performed to allow for the estimation of uncertainty in the predictions. It was observed that an increase in the number of predictors resulted in improved predictions by PTFs. It was also observed that the predictions by the SVM-based PTFs showed considerable improvement over ROSETTA.

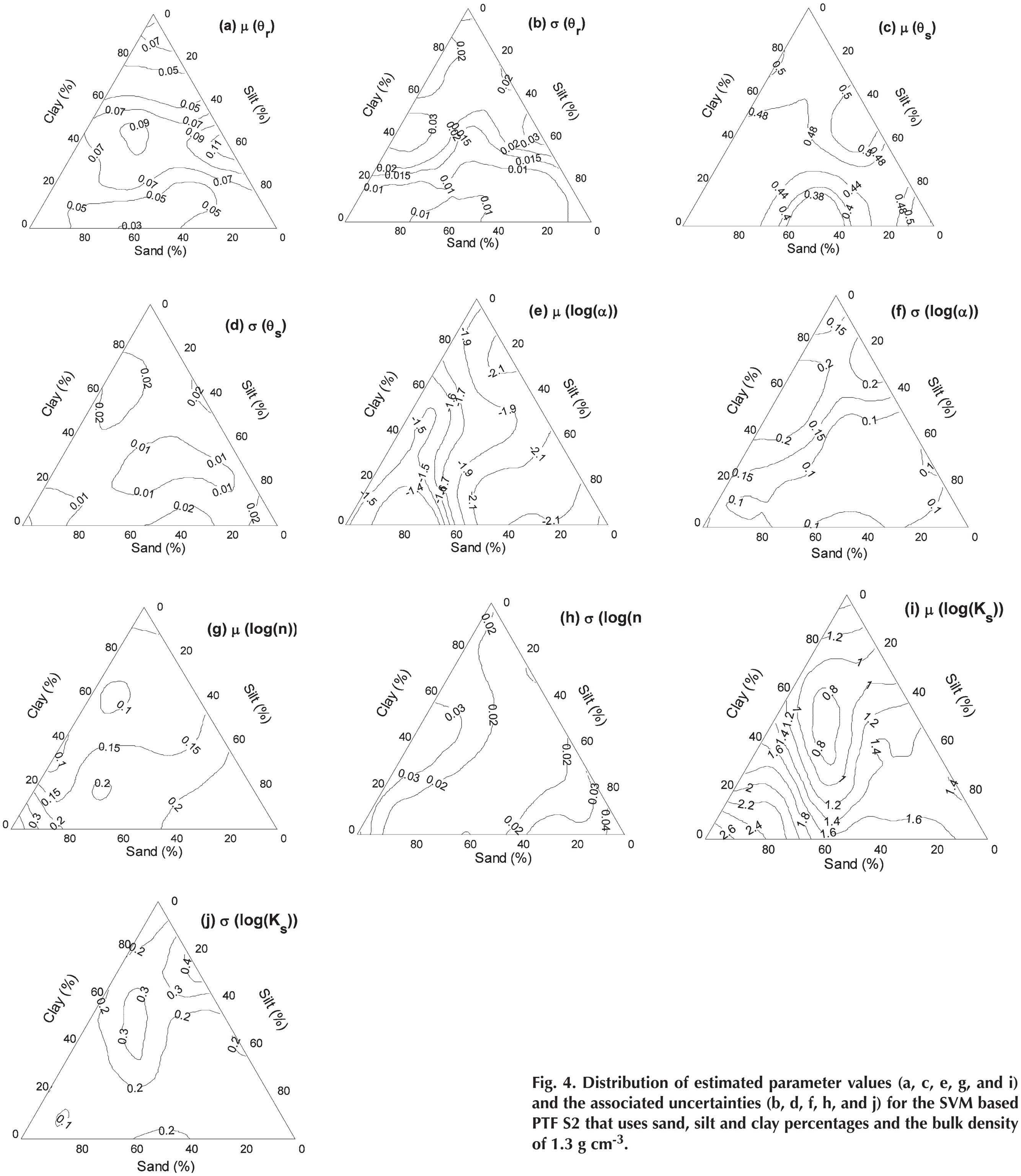

Fig. 4. Distribution of estimated parameter values (a, c, e, g, and i) and the associated uncertainties (b, d, f, h, and j) for the SVM based PTF S2 that uses sand, silt and clay percentages and the bulk density of 1.3 g cm^{-3}.

However, we note that large uncertainties in PTFs' estimates remain due to the lack of data for some textural classes (such as silt) in the training. This limitation underscores the need for gathering more information, especially for fine-textured classes (see Table 1). One may also expect that the performance of PTFs would improve if additional information, such as the organic matter content, soil structure and chemical properties, was available. However, as noted by Schaap et al. (2001), care should be taken to avoid predictors that are difficult to measure or are not commonly measured because this may put the very idea of PTFs in jeopardy.

REFERENCES

Arya, L.M., and J.F. Paris. 1981. A physicoempirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data. Soil Sci. Soc. Am. J. 45:1023–1030.

Asefa, T., M. Kemblowski, G. Urroz, M. McKee, and A. Khalil. 2004. Support vectors-based groundwater head observation networks design. Water Resour. Res. 40(11): Doi: 10.1029/2004WR003304.

Asefa, T., M. Kemblowski, U. Lall, and G. Urroz. 2005. Support vector machines for nonlinear state space reconstruction: Application to the Great Salt Lake time series. Water Resour. Res. 41(12): Doi: 10.1029/2004WR003785.

Batjes, N.H. 1996. Development of a world data set of soil water retention properties using pedotransfer rules. Geoderma 71:31–52.

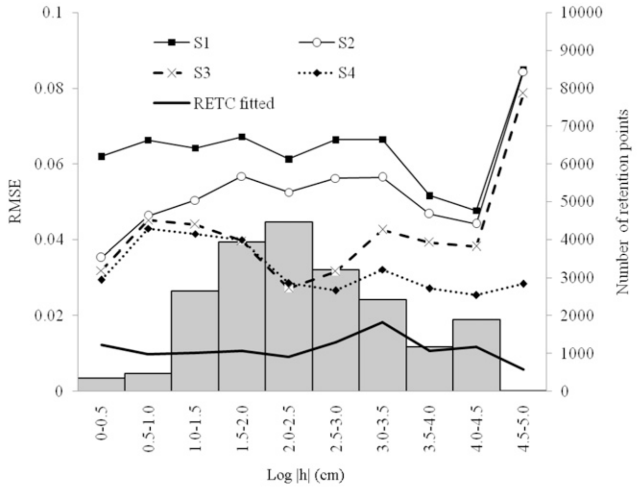

Fig. 5. RMSE for water contents predicted by the SVM-based PTFs as a function of the pressure head. RMSE for water contents estimated using the fitted parameter values is also shown. Bars show a number of retention points for different pressure head intervals.

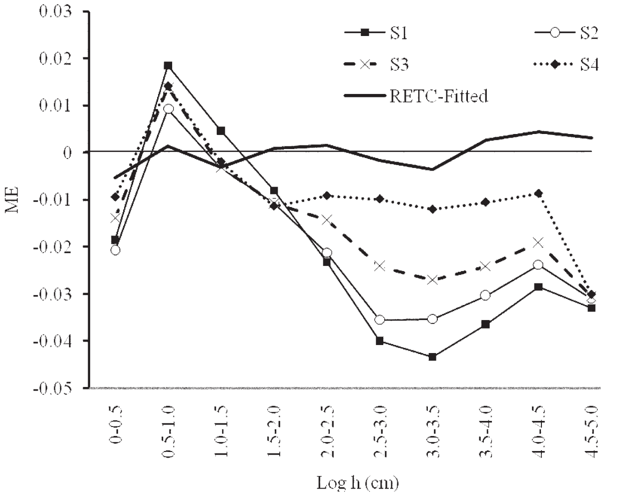

Fig. 6. Variation in Mean Error (ME) between the measured water contents and those predicted by SVM-based PTFs (S1-S4) as a function of the pressure head (cm). The ME of the direct fit to water retention data using the RETC program is also shown.

Bouma, J., and J.A.J. van Lanen. 1987. Transfer functions and threshold values: From soil characteristics to land qualities. p. 106–110. In K.J. Beek et al (ed.) Quantified land evaluation. International Institute Aerospace Surv. Earth Sci. ITC publ. 6. Enschede, the Netherlands.

Briggs, L.J., and H.L. Shantz. 1912. The wilting coefficient and its indirect measurement. Bot. Gaz. 53:20–37.

Carsel, R.F., and R.S. Parrish. 1988. Developing joint probability distributions of soil water retention characteristics. Water Resour. Res. 24:755–769.

Dibike, B.Y., S. Velickov, D. Solomatine, and B.M. Abbot. 2001. Model induction with support vector machines: Introduction and applications. J. Comput. Civ. Eng. 15:208–216.

Efron, B., and R.J. Tibshirani. 1993. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability. Chapman and Hall, New York.

Hastie, T., R.J. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: Data Mining, Inference and Prediction. Springer, New York.

Haverkamp, R., and J.-Y. Parlange. 1986. Predicting the water retention curve from particle-size distribution, 1. Sandy soils without organic matter. Soil Sci. 142:325–339.

Hopmans, J.W., D.R. Nielsen, and K.L. Bristow. 2002. How useful are small-scale soil hydraulic property measurements for large-scale vadose zone modeling. p. 247–258. *In* D. Smiles et al (ed.) Heat and mass transfer in the natural environment. The Philip Volume. Geophysical Monogr. Ser. 129. AGU, Washington, DC.

Kanevski, M., and M. Maignan. 2004. Analysis and modelling of spatial environmental data. Marcel Dekker, Berks, UK.

Leij, F.J., W.J. Alves, M.Th. van Genuchten, and J.R. Williams. 1996. The UNSODA-Unsaturated Soil Hydraulic Database. User's manual Version 1.0. Rep. EPA/600/R-96/095. National Risk Management Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH.

Liong, S.Y., and C. Sivapragasam. 2000. Flood stage forecasting with SVM. J. Am. Water Resour. Assoc. 38:173–186.

MacKay, D.J.C. 1992. Bayesian interpolation. Neural Comput. 4:415–447.

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.

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

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

Nemes, A., W.J. Rawls, Ya.A. Pachepsky, and M.Th. van Genuchten. 2006b. Sensitivity analysis of the Nonparametric Nearest Neighbor Technique to estimate soil water retention. Vadose Zone J. 5:1222–1235.

Nimmo, J.R., W.N. Herkelrath, and A.M. Laguna Luna. 2007. Physically based estimation of soil water retention from textural data: General framework, new models, and streamlined existing models. Vadose Zone J. 6:766–773.

Pachepsky, Ya.A., D.J. Timlin, and G. Varallyay. 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci. Soc. Am. J. 60:727–733.

Pachepsky, Ya.A., and W.J. Rawls. 1999. Accuracy and reliability of pedotransfer as affected by grouping soils. Soil Sci. Soc. Am. J. 63:1748–1757.

Rawls, W.J., and D.L. Brakensiek. 1985. Prediction of soil water properties for hydrologic modelling. p. 293–299. In E. Jones and T.J. Ward ed. Proc. of Symp. ASCE, Denver, CO, 30 April–2 May 1985. ASCE, New York.

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

Schaap, M.G. 2004. Accuracy and uncertainty in PTF predictions. p. 33–46. In Y. Pachepsky and W.J. Rawls (ed.) Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam, The Netherlands.

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.

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. 60:1732–1741.

Tartakovsky, D.M., and B.E. Wohlberg. 2004. Delineation of geologic facies with statistical learning theory. Geophys. Res. Lett. 31(18):L18502 10.1029/2004GL020864.

Tyler, S.W., and S.W. Wheatcraft. 1989. Application of fractal mathematics to soil water retention estimation. Soil Sci. Soc. Am. J. 53:987–996.

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.

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

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

Wohlberg, B.E., D.M. Tartakovsky, and A. Guadagnini. 2006. Subsurface characterization with support vector machines. IEEE Trans. Geosci. Rem. Sens. 44:47–57.

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.

Wösten, J.H.M., P.A. Finke, and M.J.W. Jansen. 1995. Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. Geoderma 66:227–237.

Wösten, J.H.M., Y.A. Pachepsky, and W.J. Rawls. 2001. Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251:123–150.

Ye, M., M.G. Schaap, R. Khaleel, and J. Zhu. 2007. Simulation of Field Injection Experiments in a Layered Formation Using Geostatistical Methods and Artificial Neural Network. Water Resour. Res. 43:W07413 10.1029/2006WR005030.