Received: 14 January 2020 | Revised: 21 November 2020 | Accepted: 30 November 2020

DOI: 10.1111/ejss.13077

SPECIAL ISSUE ARTICLE

European Journal of Soil Science WILEY

Development and functional evaluation of pedotransfer functions for soil hydraulic properties for the Zambezi River Basin

Mulenga Kalumba[1,2] | **Brecht Bamps**[1] | **Imasiku Nyambe**[3] | **Stefaan Dondeyne**[4] | **Jos Van Orshoven**[1]

[1]Department of Earth and Environmental Sciences, University of Leuven, Leuven, Belgium

[2]Department of Agricultural Engineering, School of Engineering, The University of Zambia, Lusaka, Zambia

[3]Department of Geology, School of Mines, The University of Zambia, Lusaka, Zambia

[4]Department of Geography, Ghent University, Ghent, Belgium

Correspondence
Mulenga Kalumba, Department of Earth and Environmental Sciences, University of Leuven, Celestijnenlaan 200E, Leuven, Belgium.
Email: mulengakalumba@yahoo.com; mulenga.kalumba@kuleuven.be

Funding information
Decision Analytic Framework 'DAFNE' EU H2020, Grant/Award Number: 690268

Abstract

Water retention and saturated hydraulic conductivity are soil properties that are key determinants in crop growth and hydrological modelling. They are commonly estimated from basic soil characteristics such as bulk density, organic carbon content and texture by means of pedotransfer functions (PTFs). In order to assess and compare the inherent performance and the functional applicability in the Zambezi River Basin (ZRB) of the widely used Saxton & Rawls PTFs and a set of newly developed PTFs, we compiled measurements of water retention at pF0.0, 1.0, 2.0, 2.8, 3.4 and 4.2 and of saturated hydraulic conductivity (Ksat) on 631 soil samples throughout the ZRB. A total of 329 of the samples were related to 55 soil profiles available in the Africa Soil Profile database, whereas our own field campaign carried out in a 2,426-km^2 subbasin of the ZRB provided the remaining 302 samples related to 119 soil profiles. Apart from evaluating the Saxton & Rawls PTFs, we developed multiple linear regression (MLR) PTFs, and PTFs derived by three machine learning (ML) models: artificial neural network (ANN), random forest (RF) and support vector machine (SVM). All PTFs were first evaluated based on a comparison of the estimated and measured property values by means of R^2, mean absolute error (MAE) and root mean squared error (RMSE). For the ensemble of MLR-PTF and ML-PTFs, the R^2 of the six water content variables and the Ksat ranged from 0.55 to 0.85, whereas for the Saxton & Rawls PTFs the range was between 0.10 and 0.50. Secondly, all PTFs were subjected to a functional evaluation using the Food and Agriculture Organization (FAO) AquaCrop crop growth model. Dry season irrigation requirements for maize as computed by AquaCrop with measured versus estimated soil hydraulic properties revealed that ANN-PTFs provide AquaCrop outputs that come closest to AquaCrop outputs generated with measured soil hydraulic properties. This study shows the

importance of performing functional evaluation of pedotransfer functions before their widespread application.

Highlights

- Developed machine learning and multiple linear regression pedotransfer functions (PTFs).
- The Saxton & Rawls PTFs are not recommended for use in the Zambezi River Basin.
- PTFs were functionally evaluated through use of estimated soil hydraulic properties in AquaCrop.
- More accurate PTFs have better functional performance, although differences are small.

KEYWORDS

artificial neural network, machine learning, multiple linear regression, random forest, saturated hydraulic conductivity, support vector machine

1 | INTRODUCTION

Soil hydraulic properties such as soil water retention characteristics (SWRC) and saturated hydraulic conductivity (Ksat) are crucial for crop growth and hydrological modelling. As such data are seldom measured, they often need to be estimated using pedotransfer functions (PTFs) that rely on basic and more easily available soil characteristics such as granulometry (sand, silt, clay and coarse fractions), bulk density, organic matter content and pH. Scanning through the International Soil Reference and Information Centre (ISRIC) Africa Soil Profiles Database (Batjes, Ribeiro, & Van Oostrum, 2019; Leenaars, van Oostrum, & Gonzalez, 2014) for the Zambezi River Basin (ZRB), we observed 1,481 legacy soil profile sites with measured basic soil characteristics for various depth layers or horizons. However there are no data on Ksat and out of these 1,481 legacy soil profiles, only 55 have water retention measurements of SWRC at two matric potentials: field capacity (pF2.0) and wilting point (pF4.2). In the absence of measured soil hydraulic properties but with the capability to sufficiently accurately estimate them with PTFs, digital soil mapping (DSM) models (McBratney, Mendonça Santos, & Minasny, 2003) can be used for estimating them at the basin scale, as, for example, for the whole ZRB. Digital maps of soil hydraulic properties can then be used as input in spatial explicit crop-water and hydrologic models.

Pedotransfer functions for soil hydraulic properties have been developed since the early 1980s (Gupta & Larson, 1979; Maclean & Yager, 1970; Rawls, Brakensiek, & Saxton, 1982; Saxton, Rawls, Romberger, & Papendick, 1986; van Genuchten, 1980). In 1996, Timlin, Pachepsky, Acock, and Whisler (1996) published a review of 49 PTFs, whereas more recently Botula, Van Ranst, and Cornelis (2014) also reviewed 35 PTFs for predicting water retention of soils in the humid tropics. The latter authors observed that 26% of the water-retention PTFs for tropical soils were developed for soils in Brazil, 26% for soils in India, 11% for soils in other countries in tropical America, and 11% for soils in Africa. Therefore, few studies have developed PTFs for soil hydraulic properties using machine learning (ML) techniques. The ones we found are all outside the Zambezi River Basin (ZRB) and include artificial neural network (ANN) (Haghverdi, Cornelis, & Ghahraman, 2012; Minasny, Mcbratney, & Bristow, 1999), random forest (RF) (Sequeira, Wills, Seybold, & West, 2014; Szabó et al., 2019) and support vector machine (SVM) (Lamorski, Pachepsky, Sławiński, & Walczak, 2008; Nguyen et al., 2017).

When PTFs are used, the variability of the basic soil properties is directly translated into variations in soil hydraulic properties and subsequently to variations in simulated functional soil behaviour (Pringle, Romano, Minasny, Chirico, & Lark, 2007; Wösten, Pachepsky, & Rawls, 2001). In addition to the effect of soil variability, the estimation error of a pedotransfer function itself results in variations in the predicted soil hydraulic properties (Araya & Ghezzehei, 2019; McNeill, Lilburne, Carrick, Webb, & Cuthill, 2018), which will also be transferred to model outputs in, for example, crop growth modelling simulation studies. Functional evaluation of pedotransfer functions dealing with soil hydraulic properties was addressed by, among others, Cresswell & Paydar (2000), Espino, Mallants, Vanclooster, & Feyen (1996), Li, Chen,

White, Zhu, & Zhang (2007), Nemes, Schaap, & Wösten (2003) and Wösten, Finke, & Jansen (1995). They evaluated the outputs of pedotransfer functions as inputs in various mechanistic models. This was also done by Timlin et al. (1996), who evaluated the outputs of pedotransfer functions as inputs into the GLYCIM crop model for four locations in Colorado (USA). They found that simulated crop yields were affected more by the PTF for estimating the water retention curve than by the PTF for estimating the saturated hydraulic conductivity (Ksat). This finding was later confirmed by Gijsman, Jagtap, and Jones (2003), who also evaluated eight PTFs for estimating the soil hydraulic properties and found that simulated soybean yield varied greatly as a consequence of the PTFs that were used to estimate the water retention curves. However, studies that compare the performance of such PTFs in terms of agreement with measured soil hydraulic properties or in terms of agreement between the output (such as irrigation water requirements) of a model fed with measured versus PTF-estimated properties are even more scarce in the ZRB. The only ones we came across include those of Gijsman et al. (2003) and Timlin et al. (1996), which are also outside the ZRB and are based on multiple linear regression (MLR) and not machine learning (ML). Moreover, in none of these studies were the irrigation water requirements used as the functional criterion to evaluate the PTFs. Therefore, in this study we performed a similar functional evaluation by using the Food and Agriculture Organization (FAO) AquaCrop crop model to investigate the relative effect of the selected PTFs for estimating water content and Ksat-values on the irrigation water requirements as calculated by AquaCrop crop model.

Despite PTFs of the MLR type being successfully applied in a wide variety of studies related to agricultural hydrology and water management, Gijsman et al. (2003) and Wassar, Gandolfi, Rienzner, Chiaradia, and Bernardoni (2016) argued that it is difficult to recommend a particular PTF due to big discrepancies between the methods used in measuring the water retention of the soil samples on which the PTFs are based. In addition, according to Minasny et al. (1999), the performance of a PTF may vary with pedological origin of the soil and location on which it was developed, and subsequently may not be directly transferable elsewhere. Furthermore, Carsel and Parrish (1988) suggest that kaolinitic tropical soils often used for agriculture usually have clay contents ranging from 60% to 90%, whereas in temperate climates, soils with more than 60% of clay are considered as heavy clays with very low saturated hydraulic conductivity (Ksat) and regarded as non-agricultural soils. This was also affirmed by Corréa (1984), Hodnett & Tomasella (2002), Minasny & Hartemink (2011), Tomasella & Hodnett (1996) and van den Berg, Klamt, van Reeuwijk, & Sombroek (1997), suggesting that kaolinitic tropical soils tend to exhibit lower bulk densities of about 0.7–1.2 g/cm^3, higher saturated hydraulic conductivity values, usually 10–1,000 mm/hr, and lower available water capacity (AWC) of about 70 mm/m when compared with typical temperate clayey soils. Their investigations highlight the importance of being cautious when selecting PTFs for a particular application.

The aim of this study was to assess, compare and functionally evaluate the performance of the broadly used PTFs of Saxton and Rawls (2006) and four newly developed PTFs to estimate saturated hydraulic conductivity and water retention at agronomically relevant matric potentials over the vast territory of the Zambezi River Basin, which has relatively scarce data. The specific objectives were: (a) to assess and compare the performance of different models for deriving PTFs where we particularly wanted to compare "classical/conventional" models relying on MLR with machine learning (ML) models as these are becoming more popular; and (b) to compare and evaluate the functional repercussions of the output obtained by these PTFs when used in the crop-water model AquaCrop (Raes, Steduto, Hsiao, & Fereres, 2009; Steduto, Hsiao, Raes, & Fereres, 2009).

1.1 | Multiple linear regression and PTFs

Multiple linear regression (MLR) requires that there is a linear relation between the dependent and the independent variables, and the independent variables are independent from each other. Furthermore, the model residuals should be normally distributed and the variance of error terms similar across the values of the independent variables. The most popular PTFs of the MLR-type are the ones proposed by Saxton and Rawls (2006). These PTFs have been promoted for use in crop growth models across the globe (e.g., by the FAO through its AquaCrop crop growth and yield prediction model) (Raes et al., 2009; Steduto et al., 2009), despite the fact that they are based on measured data for North American soils. The PTFs of Saxton and Rawls (2006) allow estimating soil moisture content of a saturated soil (pF0.0) at field capacity (pF2.0) and at wilting point (pF4.2). They also encompass a PTF to estimate the saturated hydraulic conductivity (Ksat).

1.2 | Machine learning and PTFs

In contrast to multiple linear regression (MLR), machine learning (ML) techniques do not impose stringent

assumptions about the statistical characteristics of the input data; hence they provide flexible application potential (Hastie, Tibshirani, & Friedman, 2009). These techniques have become more popular with the wider availability of powerful computing capacity and open access software such as R and in proprietary software (R Core Team, 2018), and so ML techniques do not need much greater programming skills compared with using MLR. What is different is that in contrast to MLR, ML techniques are of the black box type, meaning that extra procedures must be used to evaluate the importance of the considered predictors for the estimates obtained. ML is an umbrella term, encompassing, among others, artificial neural networks (ANNs), tree-based models such as random forest (RF) and support vector machines (SVMs).

1.3 | ANNs

In artificial neural networks (ANNs), the feed-forward back propagation is a network that is not recursive; that means neurons in one layer are only connected to neurons in the next layer, and they do not form a cycle, such that signals travel only in one direction towards the output layer (Pham, Kim, Yoon, & Choi, 2019; Sharma, Sandooja, & Yadav, 2013). In the context of regression analysis, artificial neural network models have been used as a special class of PTFs using feed-forward back propagation or radial basis functions to approximate any continuous (non-linear) function (Van Looy et al., 2017). According to Zhang and Schaap (2017), a typical feed-forward ANN contains an input layer, one or more hidden layers and an output layer, where the neurons in the hidden layer extract useful information from the input layer and utilize it to determine the output in the output layer. The number of neurons in the hidden layer is determined empirically and/or co-determined by the quality of the calibration and validation datasets.

1.4 | RF

Random forest (RF) is among a popular set of tree-based ensemble machine learning models, which are highly data adaptive and able to account for correlations as well as interactions in explanatory variables, thus making them particularly appealing for estimating soil hydraulic properties (Touw et al., 2012; Tyralis, Papacharalampous, & Langousis, 2019). RF was developed by Breiman (2001), who combined the bagging method (Breiman, 1996) with the random variable selection. In essence, bagging or bootstrap aggregation is an ensemble learning method (Sagi & Rokach, 2018) that generates a bootstrap sample from an original training dataset and trains a model using the generated sample. This procedure is repeated (*ntree*) times so that the bagging's prediction is the average of the predictions of the (*ntree*) trained models. This way, bagging reduces the variance of the prediction function, although it requires that unbiased models are efficient (Hastie et al., 2009).

1.5 | SVM

Support vector machines (SVMs) are based on simple ideas that originated in statistical learning theories (Karatzoglou, Meyer, & Hornik, 2006). SVM regression is based on the generalized regression formulation, where the explanatory variables are first mapped onto an m-dimensional space using some fixed (non-linear) mapping, and then a linear regression model relating the explanatory and response variables is constructed (Twarakavi, Šimůnek, & Schaap, 2009). The goal of SVM is to create a smooth boundary, called a hyperplane, which leads to fairly homogeneous partitions of observations on either side of the plane (Lantz, 2013). Support vector machines gained popularity in many fields that were traditionally dominated by ANNs. SVMs have the advantage of being more robust and efficient than ANNs as they allow neglecting small errors, making the regression sparse and so avoiding local minimum issues. They also require fewer input variables/data/calibration points (Lamorski et al., 2008; Valyon & Horváth, 2005). In this way, SVMs are less susceptible than ANNs to overfitting (Lamorski et al., 2008; Yi Lin, Cheng, & Wing Chau, 2007). The simplicity comes from the fact that an SVM applies a simple linear method to the data but in a high-dimensional feature space that is non-linearly related to the input space and can be imagined as a surface that defines a boundary between various points of data that represent examples plotted in multidimensional space according to their feature values.

2 | MATERIALS AND METHODS

2.1 | Soil data

The Zambezi River Basin (ZRB), covering approximately 1,600,000 km^2, is located between 8–20° S latitude and 17–36° E longitude in southern Africa (Figure 1). The soil data was obtained from the Africa Soil Profiles Database (Batjes et al., 2019; Leenaars et al., 2014). For the ZRB, this database contains data of 1,481 georeferenced legacy soil profiles encompassing 5,184 soil horizons or layers for which measured elementary soil characteristics are

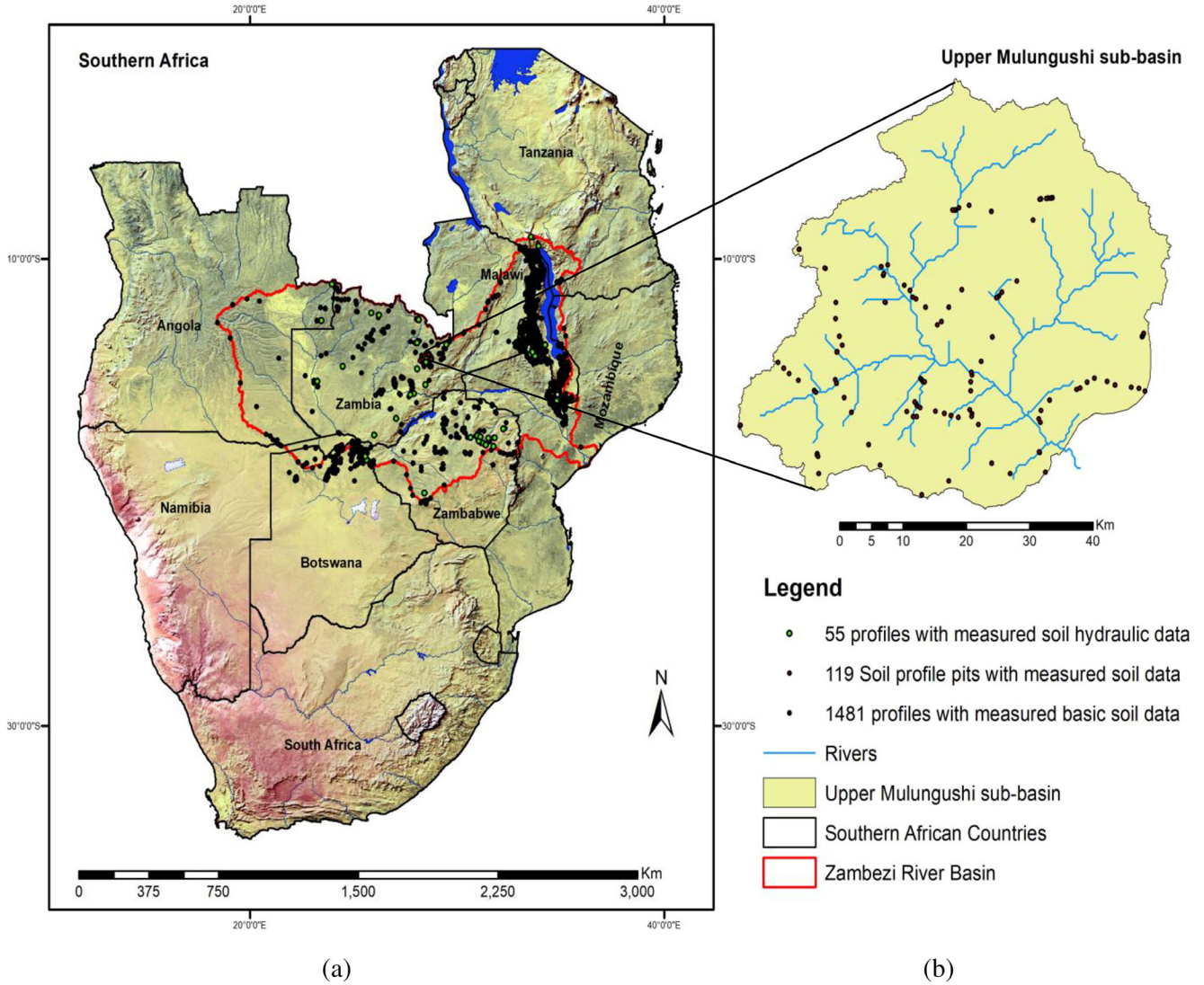

FIGURE 1 Distribution of the soil profiles (a) from the African soil profile database and (b) own sampling in the upper-Mulungushi subbasin in central Zambia [Color figure can be viewed at wileyonlinelibrary.com]

available: soil textural composition (sand, silt and clay fractions), soil organic carbon content (OC %), pH in 1:5 H_2O, bulk density (BD, g/cm^3), depth of the upper and lower boundaries of the horizon or layer (Dep, cm). Also, topographic elevation (ELE, m) of the profile site is available. Among these 1,481 profiles, there are only 55 profiles associated with 329 soil horizons or layers for which water contents at two matric potentials, pF2.0 and pF4.2, are recorded.

In the smaller 2,426 km^2 Upper Mulungushi subbasin (UMB) of the ZRB (Figure 1), we collected 302 bulk samples from 119 soil profiles at depth layers of 30, 60 and 100 cm. These were analysed for soil textural composition (sand, silt and clay fractions), stoniness (STON %), soil organic carbon content (SOC %), nitrogen content (N %), pH in 1:5 H_2O and electrical conductivity (EC). From the same profiles and at the same depth layers, we also took undisturbed core samples with Kopecky rings (100 cm^3) for measuring saturated soil hydraulic conductivity (Ksat), and the water contents at six matric potentials: pF0.0 (saturation), pF1.0, pF2.0 (field capacity), pF2.8, pF3.4 and pF4.2 (wilting point). The Ksat was measured in the laboratory by placing the Kopecky rings with undisturbed soil samples in a constant head permeameter apparatus. Table 1 shows the summary statistics of all the soil data in the UMB and ZRB used in this study, with the last column indicating that the PTFs for water content at pF2.0 and 4.2 are based on the 631 measurements from the African Soil Profile (AfSP), whereas the PTFs for the five other variables of interest are based on 302 measurements from the UMB.

TABLE 1 Summary statistics of all the 631 soil samples, 329 of which are from the African soil profile (AfSP) and spread throughout the Zambezi River basin (Batjes et al., 2019; Leenaars et al., 2014), and 302 are own data samples from the upper-Mulungushi River basin (UMB), a sub-basin of the Zambezi River basin

Explanatory variables							
	Min	**1 st Q**	**Median**	**Mean**	**3 rd Q**	**Max**	**Source and number of samples**
Dep (cm)	6	30	60	75	100	217	AfSP 631
ELE (m)	81	1,146	1,173	1,179	1,222	1924	AfSP 631
SAND (%)	3.20	33.10	49.60	50.80	67.40	97.30	AfSP 631
CLAY (%)	1.50	13.30	24.60	27.30	39.50	84.30	AfSP 631
SILT (%)	4.70	11.00	19.30	21.90	30.40	78.60	AfSP 631
OM (%)	0.039	0.041	0.170	0.337	0.421	6.854	AfSP 631
N (%)	0.000	0.025	0.034	0.040	0.046	0.366	UMB 302
EC (ds/m)	0.001	0.008	0.012	0.016	0.018	0.090	UMB 302
pH	4.98	5.60	6.06	5.97	6.43	8.50	AfSP 631
STON (%)	0.00	0.00	0.00	6.20	0.00	85.00	UMB 302
BD (g/cm^3)	0.83	1.46	1.52	1.53	1.59	2.04	AfSP 631
Por (%)	23.00	40.10	42.30	42.10	45.00	68.60	AfSP 631
Response variables							
Ksat (mm/hr)	0.10	22.40	59.20	107.60	130.80	509.40	UMB 302
pF0.0 (Vol %)	21.20	34.40	38.70	38.50	42.20	66.10	UMB 302
pF1.0 (Vol %)	19.00	32.00	36.40	36.10	39.60	63.70	UMB 302
pF2.0 (Vol %)	3.60	14.40	20.70	20.40	25.70	54.30	AfSP 631
pF2.8 (Vol %)	2.60	10.20	15.40	15.60	18.90	50.70	UMB 302
pF3.4 (Vol %)	1.00	7.90	13.10	13.20	16.20	49.50	UMB 302
pF4.2 (Vol %)	0.80	4.60	10.50	11.00	15.40	48.50	AfSP 631

Note: Minimum (Min), interquartile range (IQR) = 1st Q, 3rd Q, and maximum (Max). BD, bulk density; Dep, depth; EC, electrical conductivity; ELE, elevation; N, nitrogen content; OM, organic matter; Por, porosity; STON, stoniness.

2.2 | Variable selection and PTF development

Data preprocessing, variable selection, model training and testing were carried out by means of R software version 3.5.0 (R Core Team, 2018). Table 1 shows that we had 12 explanatory variables, namely: textural composition (sand, silt and clay fractions), stoniness (STON %), organic matter content (OM %), nitrogen content (N %), pH in 1:5 H_2O, electrical conductivity (EC), bulk density (BD, g/cm^3), porosity (Por, %), depth of upper and lower boundary (Dep, cm) and topographic elevation (ELE, m), whereas there were seven response or dependent variables, including the saturated hydraulic conductivity (Ksat), and water contents at six matric potentials: pF0.0, pF1.0, pF2.0, pF2.8, pF3.4 and pF4.2. For each response variable, potential predictors were selected from the 12 available candidates by iteratively incorporating them into 12 possible models until the Akaike information criterion (AIC) stopped decreasing while using the backward-stepwise (removes) selection AIC *R package* (Kuhn & Johnson, 2013). We therefore selected models with potential explanatory variables that had the lowest AIC number for each particular response variable.

After the potential variables were selected, the overall datasets were randomly split into a training (70%) and a test dataset (30%). For the machine learning PTFs, we first normalized both the explanatory and response variables before randomly splitting the datasets into training sets (70%) and test sets (30%). With these datasets and for each response variable, PTFs of the multiple linear regression (MLR), artificial neural network (ANN), random forest (RF) and support vector machine (SVM) types were trained and tested using the respective *R packages*. To guarantee a fair comparison with the Saxton and MLR PTFs and avoid over-fitting the machine learning PTFs, the default meta-parameters of the ML models were used, such that no model tuning was done.

The performance of all these models and of the pre-existing Saxton and Rawls (2006) PTFs was first

evaluated by comparing estimated and measured values using the coefficient of determination (R^2), the mean absolute error (MAE) and the root mean squared error (RMSE). Well-performing models had high R^2, low MAE and low RMSE. Furthermore, the predictors that were selected for each PTF during training and testing were ranked according to their order of importance relative to a particular response variable using the *relative importance function* of the random forest *R package*.

2.3 | Functional evaluation of the PTFs using the FAO AquaCrop crop model

After training and testing each PTF, we further evaluated their performance by evaluating and inter-comparing the outputs (dry season irrigation water requirements) of the AquaCrop model operated with either the field reference data (measured soil hydraulic properties) or the PTF outputs (estimated soil hydraulic properties) as input in the AquaCrop. The AquaCrop model uses a small number of explicit parameters and largely intuitive input variables, either widely used or requiring simple methods for their determination (Raes et al., 2009). The input consists of weather data time series, crop variables, soil properties and field management practices that define the environment in which a crop develops (Figure 2). In our functional evaluations, except for soil characteristics such as rootable depth, water contents at three matric potentials, pF0.0 (saturation), pF2.0 (field capacity) and pF4.2 (wilting point), and the Ksat values, all the data inputs such as weather and crop variables were kept constant for each profile site for which simulations were made by means of AquaCrop.

2.4 | Climate

We used the weather data from the Global Yield Gap and Water Productivity Atlas (GYGA) (Van Wart et al., 2015, 2013) for the Kabwe climate station at 14° S latitude and 28° E longitude in Zambia. Kabwe is a town in the Upper Mulungushi Basin. The data consisted of daily records for 1998–2012 of rainfall, minimum and maximum temperature, mean relative humidity, mean wind speed and solar radiation. The time series of daily reference evapotranspiration (ETo) that AquaCrop also requires was calculated from minimum and maximum temperature, mean relative humidity, mean wind speed and solar radiation using the ETo calculator (Raes et al., 2009).

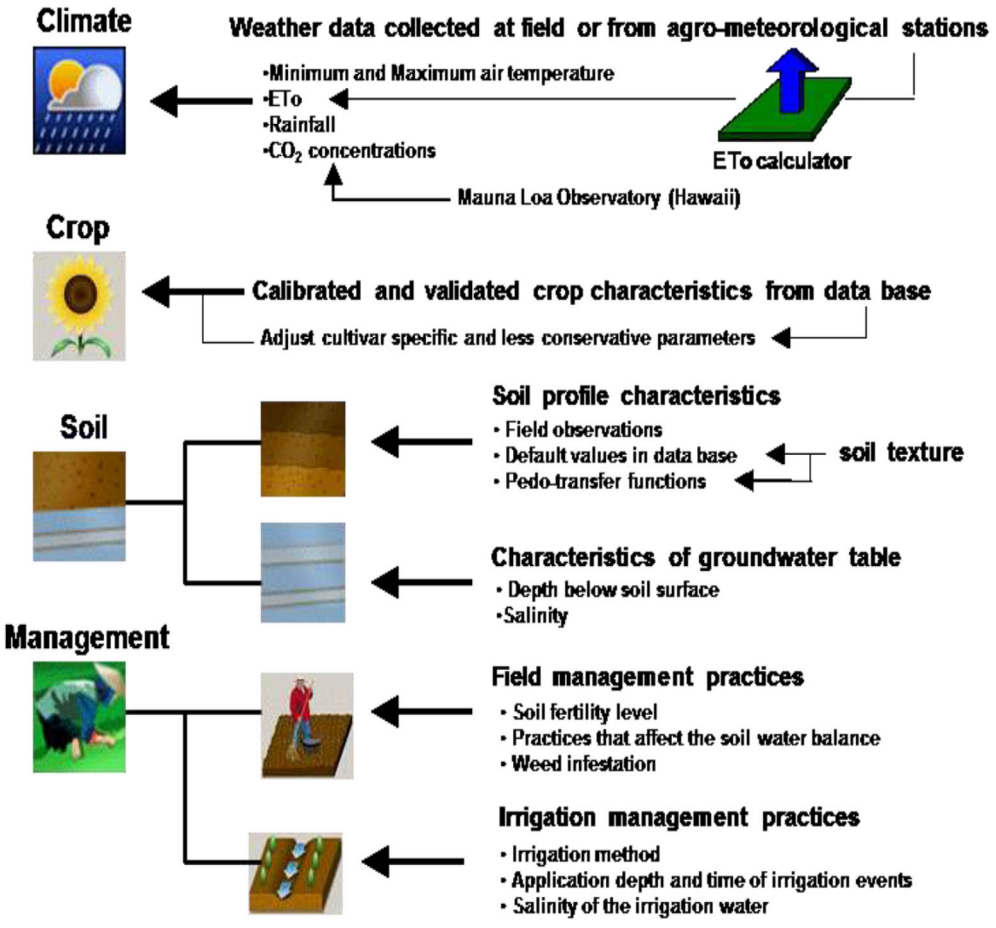

FIGURE 2 Input data for AquaCrop defining the simulation environment (Raes et al., 2009) [Color figure can be viewed at wileyonlinelibrary.com]

2.5 | Crop

We choose maize (*Zea mays* L.) for simulating the water-crop productivity as it is the most common crop grown by both smallholders and large commercial farmers throughout the ZRB. In AquaCrop, crops are characterized by a set of parameters that can easily be calibrated by the modeller to meet specific local growing conditions (Raes et al., 2009). For this study, we used the default crop parameters, such as the length of a growing period, planting density, the canopy crop development and root depth for maize. The planting date was set as April 15 for the dry season simulations because this is when farmers plant maize under irrigation after the rainy season has ended. The actual simulations started on January 1 prior to the planting date. At that time the initial soil water content was assumed to be at field capacity just for the first year of simulations throughout the soil profile.

2.6 | Soil

Based on the morphologic and chemical characteristics, the 119 soil profiles in the Upper Mulungushi Basin (UMB) were classified in Reference Soil Groups (RSGs) according to the third edition of the World Reference Base for soil resources (IUSS Working Group WRB, 2015). There were eight RSGs identified: Acrisols, Arenosols, Gleysols, Lixisols, Phaeozems, Plinthosols, Regosols and Vertisols. From the 119 soil profiles, we randomly selected one agricultural site for each of the eight RSGs for simulating the irrigation requirements for maize. The soil textural classes of each RSG selected were as presented in Table 2.

TABLE 2 The eight Reference Soil Groups (RSGs) with their textural classes representing the simulation sites used in AquaCrop simulations

RSG	Depth layers (cm)	FAO textural class
Acrisol (simulation site 1)	30	Sandy loam
	60	Clay loam
	100	Clay
Arenosol (simulation site 2)	30	Loamy sand
	60	Loamy sand
	100	Loamy sand
Gleysol (simulation site 3)	30	Silty clay
	60	Clay loam
	100	Loam
Lixisol (simulation site 4)	30	Clay loam
	60	Clay loam
	100	Clay loam
Phaeozem (simulation site 5)	30	Loam
	60	Loam
	100	Loam
Plinthosol (simulation site 6)	30	Sandy loam
	60	Sandy loam
	100	-
Regosol (simulation site 7)	30	Sandy loam
	60	Sandy loam
	100	Sandy loam
Vertisol (simulation site 8)	30	Silty clay
	60	Silty clay
	100	Silty clay

3 | RESULTS

The outputs of all the PTFs for the water contents at the six matrix potentials are expressed in fractions or percentages, whereas the Ksat values are expressed in mm/hr. The predictors sand, clay and silt content are also expressed as fractions for input in the PTFs.

3.1 | Selected potential explanatory variables

The results in Table 3 present the water content at pF0.0 and pF1.0 with eight potential explanatory variables selected from the possible 12 (Table 1) that were available for selection. The bulk density, sand fraction, nitrogen content, silt fraction and the pH were the five most important predictors out of the eight. The eight potential predictors were then used to train and test the MLR and machine learning PTFs for water content at pF0.0 and pF1.0. For estimating the water content at pF2.0, pF2.8 and pF3.4, 10 potential predictors were selected from the possible 12 variables, with clay and sand fractions, nitrogen content, silt fraction, bulk density and organic matter content as the most important predictors. The MLR and machine learning PTFs were then trained and tested using these 10 potential variables.

Nine potential variables were selected from the available 12 variables for the water content at pF4.2. At this point, clay and sand fractions, nitrogen content, organic matter content and silt fraction were the five most important predictors out of the nine potential ones. Only four variables were selected from the possible 12 candidate

TABLE 3 Potential explanatory variables selected for the multiple linear regression (MLR) and machine learning pedotransfer functions (PTFs) ranked in order of importance, with rank 1 being the most important

Selected potential explanatory variables and rank number													
pF0.0		**pF1.0**		**pF2.0**		**pF2.8**		**pF3.4**		**pF4.2**		**Ksat**	
Variable	**Rank**	**Variable**	**Rank**	**Variable**	**Rank**	**Variable**	**Rank**	**Variable**	**Rank**	**Variable**	**Rank**	**Variable**	**Rank**
BD	1	BD	1	Clay	1	Clay	1	Clay	1	Clay	1	STON	1
Sand	2	Sand	2	Sand	2	Sand	2	Sand	2	Sand	2	Sand	2
N	3	N	3	N	3	N	3	N	3	N	3	BD	3
Silt	4	Silt	4	Silt	4	OM	4	OM	4	OM	4	N	4
pH	5	pH	5	BD	5	Silt	5	Silt	5	Silt	5		
ELE	6	ELE	6	ELE	6	EC	6	EC	6	EC	6		
EC	7	EC	7	EC	7	BD	7	BD	7	ELE	7		
STON	8	STON	8	pH	8	ELE	8	ELE	8	STON	8		
				STON	9	Dep	9	Dep	9	Dep	9		
				Dep	10	STON	10	STON	10				

BD, bulk density; Dep, depth; EC, electrical conductivity; ELE, elevation; N, nitrogen content; OM, organic matter; STON, stoniness.

TABLE 4 Intercepts and coefficients for the multiple linear regression (MLR) pedotransfer functions (PTFs)

Response variables							
Predictors	**pF0.0**	**pF1.0**	**pF2.0**	**pF2.8**	**pF3.4**	**pF4.2**	**Ksat**
Dep			$3.06 \times e^{-4}$	$4.17 \times e^{-4}$	$5.31 \times e^{-4}$	$3.81 \times e^{-4}$	
ELE	$-1.17 \times e^{-4}$	$-4.83 \times e^{-5}$	$-2.21 \times e^{-4}$	$-2.43 \times e^{-4}$	$-2.30 \times e^{-4}$	$-1.79 \times e^{-4}$	
SAND	−0.073	−0.101	−1.781	−2.401	−2.347	−1.811	$1.70 \times e^{2}$
CLAY			−1.487	−2.026	−1.957	−1.430	
SILT	−0.052	−0.098	−1.712	−2.320	−2.282	−1.761	
STON	−0.024	−0.034	−0.022	−0.022	−0.023	−0.023	$4.83 \times e^{2}$
OM				0.014	0.016	0.015	
N	0.181	0.350	0.900	0.887	0.787	0.623	$-4.26 \times e^{2}$
EC	−0.360	−0.340	0.317	0.595	0.703	0.754	
pH	−0.003	−0.005	−0.010				
BD	−0.319	−0.273	−0.056	0.029	0.027		$-2.37 \times e^{2}$
Intercept	1.090	0.948	2.276	2.610	2.522	1.972	$3.86 \times e^{2}$

BD, bulk density; Dep, depth; EC, electrical conductivity; ELE, elevation; N, nitrogen content; OM, organic matter; STON, stoniness.

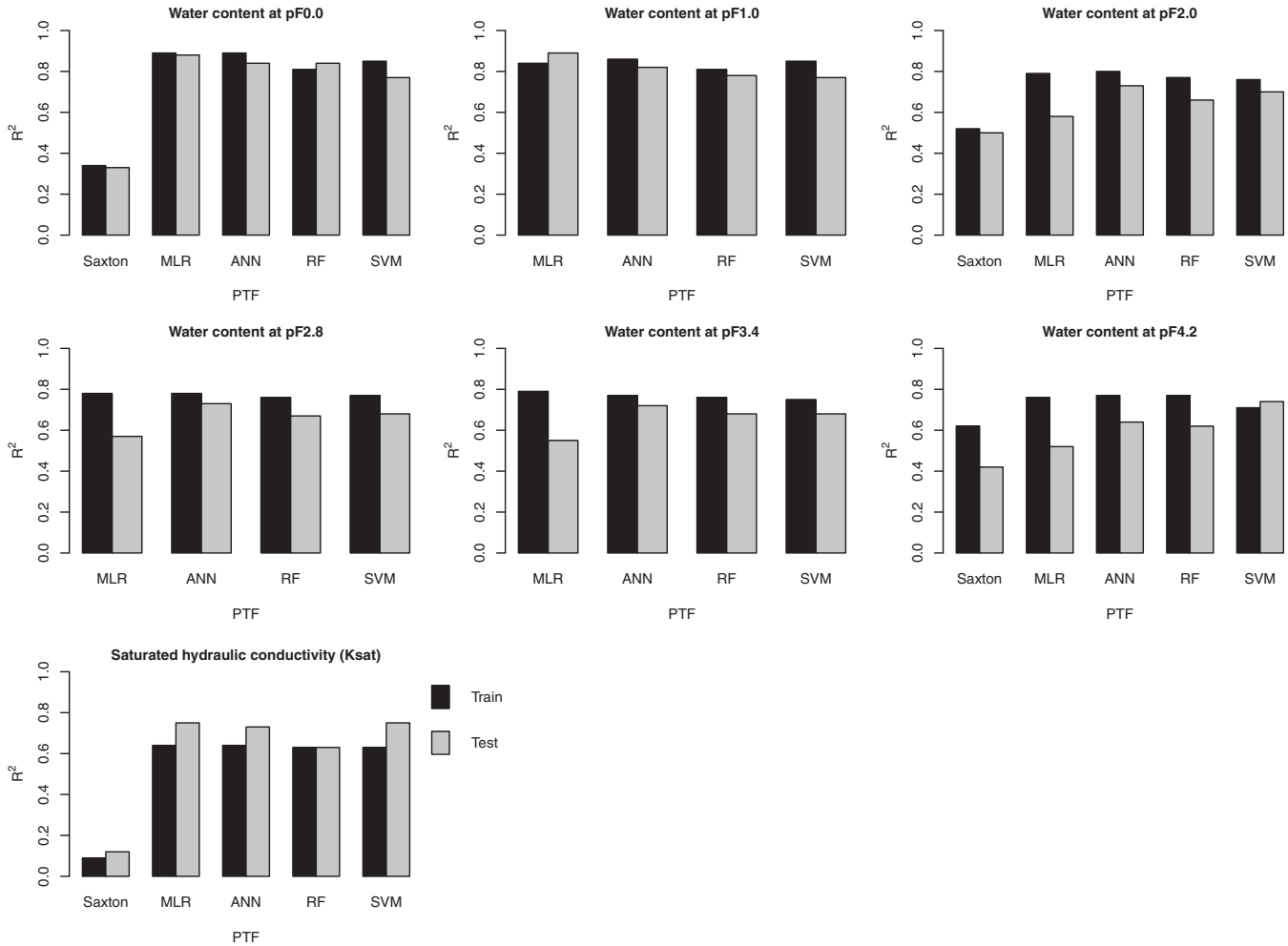

FIGURE 3 Coefficients of determination (R^2) for the pedotransfer functions (PTFs) applied to the training sets (black) and test datasets (grey). ANN, artificial neural network; MLR, multiple linear regression; RF, random forest; SVM, support vector machine

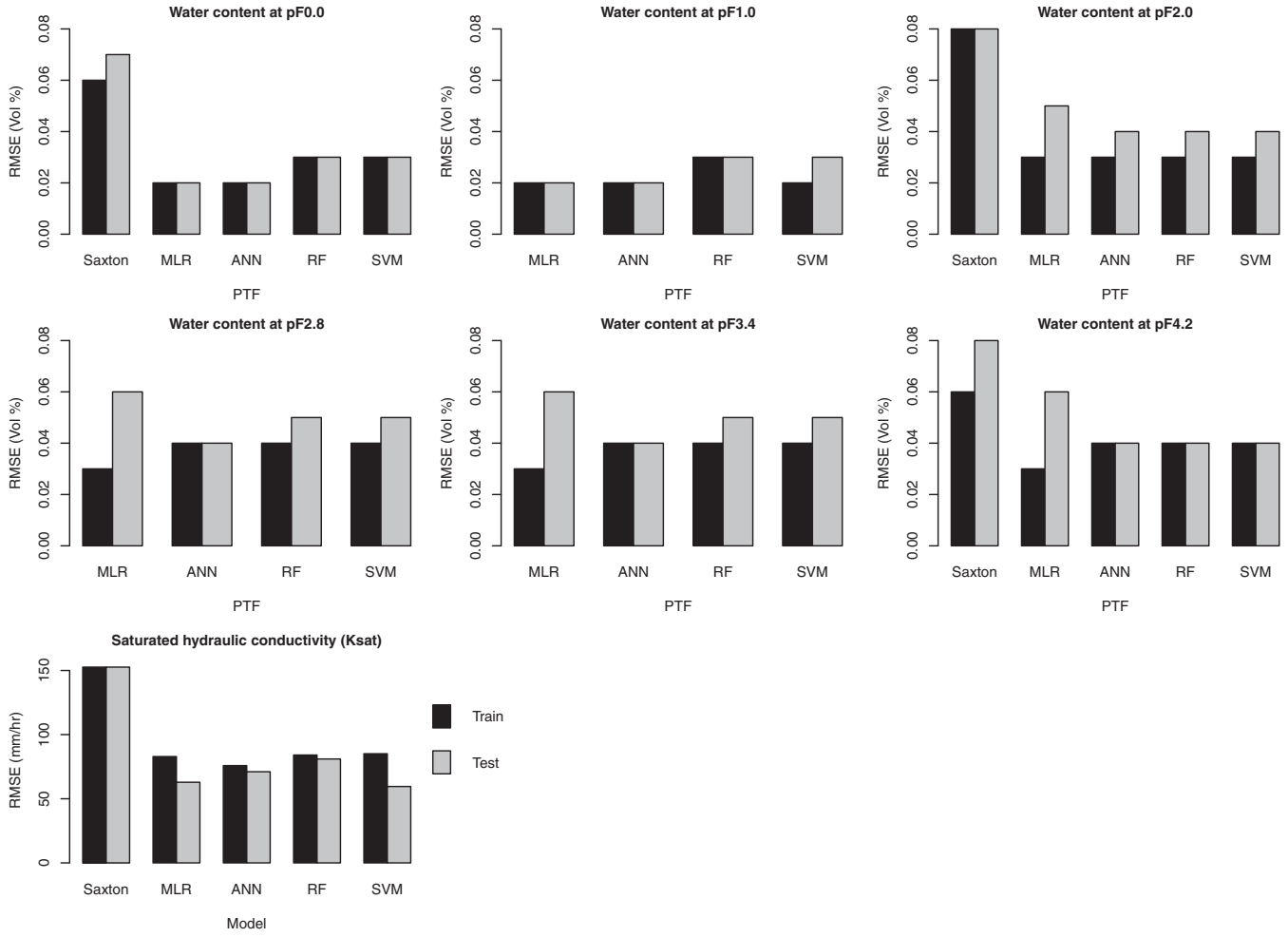

FIGURE 4 Root mean squared error (RMSE) for the PTFs applied to the training sets (black) and test datasets (grey). ANN, artificial neural network; MLR, multiple linear regression; RF, random forest; SVM, support vector machine

variables for Ksat, with stoniness and sand fraction as the two most important predictors.

In the Saxton & Rawls PTF, the sand fraction and estimated water content at pF2.0 are used as predictors for estimating water content at pF0.0. Furthermore, sand and clay fractions and organic matter content (OM), as well as their interactions (e.g., SAND * OM), are also used as predictors for water contents at pF2.0 and pF4.2. The saturated hydraulic conductivity, Ksat, is estimated using an equation also found in Saxton and Rawls (2006), with the water content at pF0.0 as the most important predictor. The intercepts and coefficients of the selected predictors for the MLR PTFs are presented in Table 4.

3.2 | Performance evaluation of the PTFs

Figures 3 to 5 present the results of the performance evaluation of the PTFs in terms of the statistical indices R^2, MAE and RMSE for both the training and test datasets. The lowest RMSE and MAE and the highest R^2 were observed for the MLR and machine learning PTFs. In Figure 3, for the water content at pF0.0, pF2.0 and pF4.2 and Ksat values, the R^2 ranged from 0.55 to 0.85. At pF1.0, pF2.8 and pF3.4, the R^2 ranged from 0.60 to 0.85 in training and test sets for the MLR and machine learning PTFs. Meanwhile, for the Saxton & Rawls PTF, the R^2 ranged from 0.10 to 0.50 for the water content at pF0.0, pF2.0, pF4.2 and Ksat values.

In Figure 4, the RMSE ranged from 0.02 to 0.06 (Vol %) and 50 to 85 (mm/hr) in training and test sets for the water content at pF0.0, pF2.0 and pF4.2 and Ksat values for the MLR and machine learning PTFs. However, for the Saxton & Rawls PTF, the RMSE ranged from 0.06 to 0.08 (Vol %) and 100 to 150.61 (mm/hr) for the water content at pF0.0, pF2.0 and pF4.2 and Ksat values, whereas for the water content at pF1.0, pF2.8 and pF3.4, the RMSE ranged from 0.02 to 0.06 (Vol %) in training and test sets for the MLR and machine learning PTFs.

Similar to the RMSE, from Figure 5, the MAE of the Saxton & Rawls PTF is much higher than the MAE of

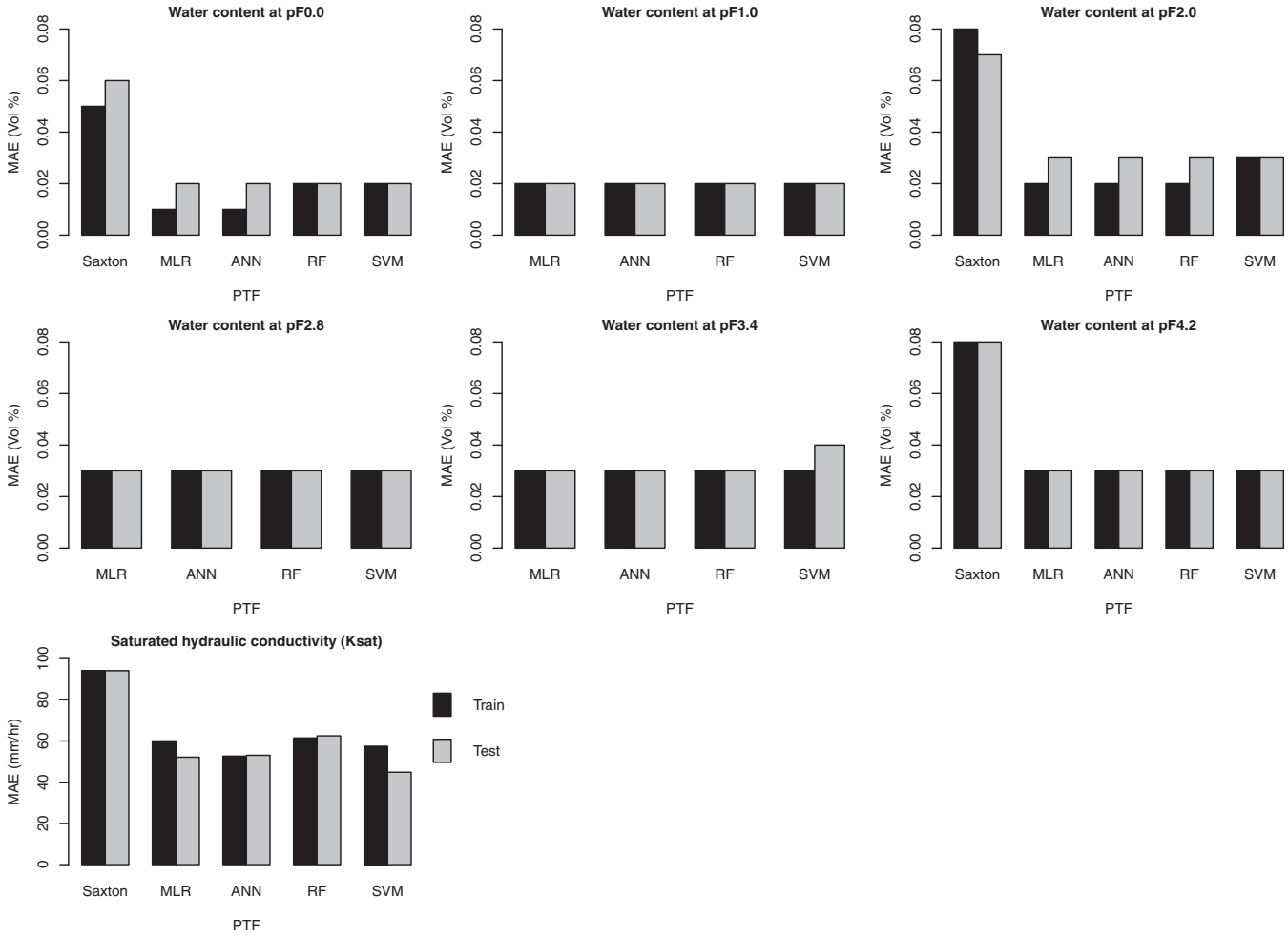

FIGURE 5 Mean absolute error (MAE) for the PTFs applied to the training sets (black) and test datasets (grey). ANN, artificial neural network; MLR, multiple linear regression; RF, random forest; SVM, support vector machine

MLR and machine learning PTFs. The differences in MAE and RMSE (Figure 4) between the MLR and machine learning PTFs were rather marginal.

3.3 | Functional evaluation of the PTFs using AquaCrop model outputs

Figure 6 displays the results as outputs of the AquaCrop model: the irrigation water requirements in the dry season for a maize crop in the simulation period (1998–2012).

In the dry season irrigation water requirements of Figure 6, we observe that on all the soil profiles except for a Gleysol soil profile with Clayic and Loamic soil textures (Table 2), the ANN followed by the MLR PTFs produced estimations of hydraulic soil properties that, when used as input in the AquaCrop model, translated into outputs of irrigation water requirements most comparable to those when measured hydraulic soil properties are used as input in the AquaCrop model.

4 | DISCUSSION

Soil hydraulic properties are key inputs for hydrological models in general and in crop-water models in particular. Traditionally, many studies have developed PTFs to estimate hydraulic soil properties from basic soil properties (e.g., bulk density, % clay, % sand and % silt) (Li et al., 2007; Schaap, Leij, & van Genuchten, 1998, 2001). Our study affirms the importance of these basic soil properties as predictors for soil water retention and saturated hydraulic conductivity (Ksat), complemented by the % stoniness (Table 3), which is highly ranked when it comes to estimating Ksat. Schaap et al. (2001) developed multiple linear regression (MLR)-based PTFs for Ksat using soil texture (% clay, % sand and % silt) among the inputs, and found an R^2 up to 0.54, whereas the R^2 of the ANN-based PTF in Schaap et al. (1998) was up to 0.57. The R^2 for the PTF in Li et al. (2007) for Ksat also went up to 0.66. We observed that the R^2 (0.55 to 0.85) in our study for the MLR and machine learning PTFs was slightly better than

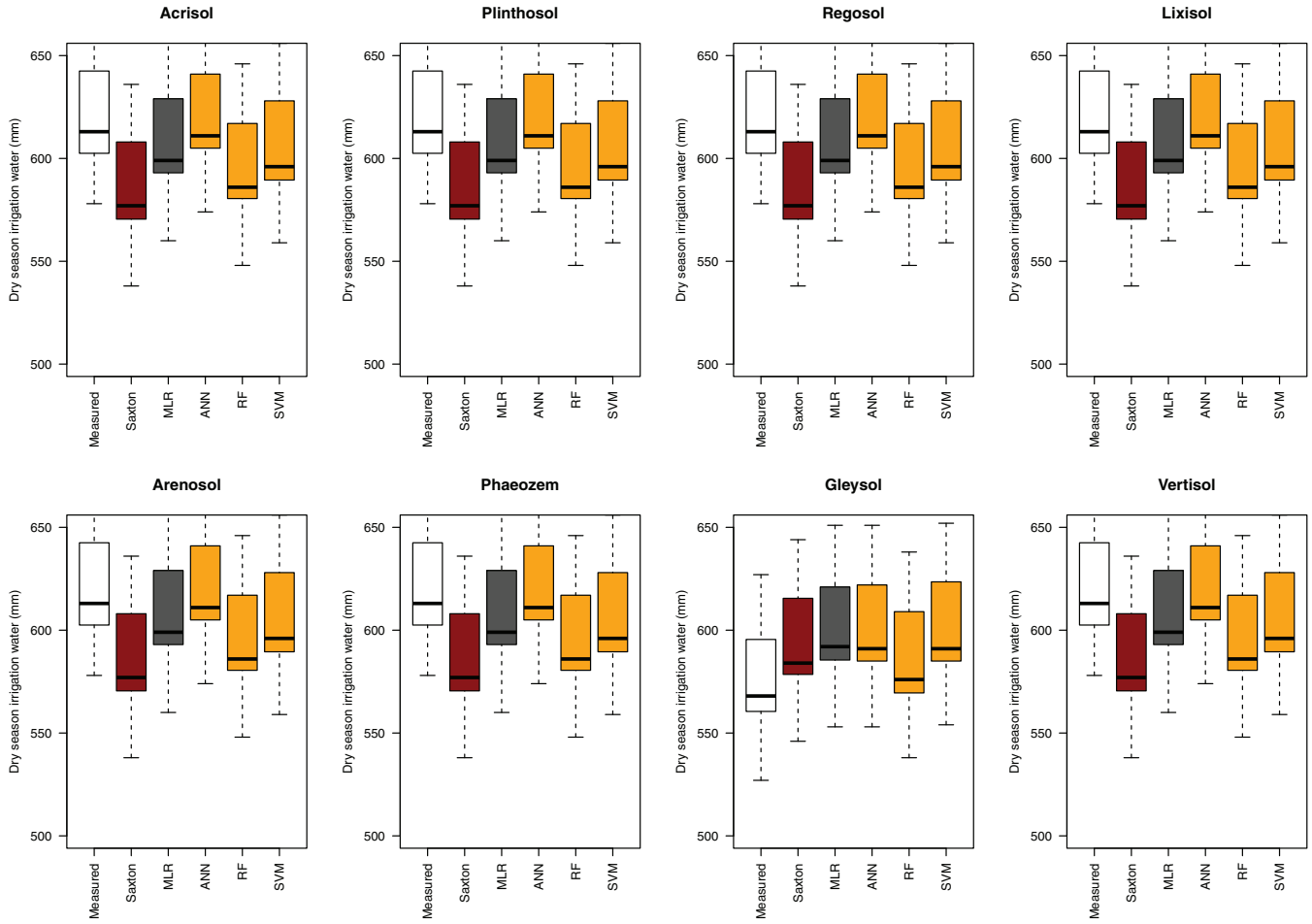

FIGURE 6 Irrigation water requirements in the dry season simulated for a maize crop for the years 1998–2012 on eight reference soil groups. The simulations are performed with AquaCrop, whereby next to the measured hydraulic properties (white), the Saxton (red), the MLR (grey) and three machine learning PTFs (orange) are used for estimating the soil hydraulic properties [Color figure can be viewed at wileyonlinelibrary.com]

in the other previous studies. For the Saxton & Rawls PTF, R^2 obtained in this study is lower than the values reported in previous studies, with a range of about 0.10 to 0.50. For Ksat, R^2 is even lower, with a range of about 0.1 to 0.12 in both the training and test processes. This rather low performance of the Saxton & Rawls PTF illustrated by the statistical indices in this study when compared to the MLR and machine learning PTFs could be attributed to the different methods that Saxton & Rawls used to measure the hydraulic soil properties, the differences in clay mineralogy of the soils that were studied, differences in land-management practices, and most probably differences between the climate of Zambia and that of the North American region from where the samples on which the Saxton & Rawls PTFs are based originate.

Through the functional evaluation of the PTFs, we observed that the performance in terms of R^2, RMSE and MAE of the Saxton & Rawls PTF was lower than that of the MLR and machine learning PTFs. In the dry season irrigation water requirements, the ANN followed by MLR PTFs performed better than the other PTFs on all the soil profiles except for a Gleysol soil profile with Clayic and Loamic soil textures, where the Saxton & Rawls PTF and the RF PTFs were slightly better. However, dry season irrigation requirements for maize as computed by AquaCrop with measured versus estimated soil hydraulic properties revealed that ANN-PTFs provide AquaCrop outputs that come closest to AquaCrop outputs (dry season irrigation requirements) generated with measured soil hydraulic properties. The AquaCrop model, when operated with measured soil hydraulic properties, indicates on average an irrigation water requirement of around 600 mm/growing season, which is in line with what farmers told us they need for growing maize in the dry season in the Kabwe (Zambia) area. This is also in line with the FAO irrigation guidelines, which suggest that for better production, a medium-matured maize crop requires between 500 to 800 mm of water depending on

the environment (FAO, 2012). This gives us confidence that we work with sufficiently realistic reference values, to which we then compare the irrigation requirements modelled using the estimated soil hydraulic characteristics. In this comparison, we used the same values of rooting depth, crop variety (physiological length of growing season) and the rainfall distribution from actual weather data for 1998–2012. We acknowledge that rooting depth, length of growing season and rainfall distribution vary through the examined sites and will affect the modelling results, but by excluding this variation from the modelling, we attempt to functionally evaluate the PTFs regardless of the other variations.

Overall this study confirms that PTFs that deliver more accurate estimates of the soil hydraulic properties return the more consistent simulation results from a crop model. Hence, this study shows the importance of performing functional evaluation of pedotransfer functions before their widespread application.

5 | CONCLUSION

This paper presented the development, assessment and functional evaluation of pedotransfer functions for estimating soil hydraulic properties in the Zambezi River Basin, from basic soil properties using MLR, ANN, RF and SVM methods. Also, the widely used pre-existing Saxton & Rawls PTFs were assessed and evaluated. Overall the Saxton & Rawls PTFs had lower R^2 and higher RMSE and MAE values than the MLR and machine learning PTFs. On the other hand, there were marginal differences in the statistical indices, R^2, RMSE and MAE, between the MLR and machine learning PTFs.

There were systematic underestimations of dry season irrigation requirements for maize as computed by AquaCrop with measured versus estimated soil hydraulic properties from all the PTFs except for the ANN PTF. On a Gleysol, the Saxton and RF PTFs performed slightly better but with some slight overestimations in the dry season irrigation water requirements. However, because the ANN PTF performed as well as the MLR and the other machine learning PTFs, with higher R^2 and lower RMSE and MAE than the Saxton & Rawls PTF, and the fact that it also performed better during functional evaluation of the dry season irrigation water requirements compared with AquaCrop model outputs, we recommend estimating the soil hydraulic properties from the available basic soil data using the ANN PTF in the Zambezi River Basin.

ACKNOWLEDGEMENT

This research was financed by an IRO PhD scholarship provided by the University of Leuven (KU Leuven) and was conducted in the framework of the Decision Analytic Framework to explore the water-energy-food Nexus, "DAFNE" EU H2020-project (grant no. 690268), and through which part of the fieldwork could be conducted. The authors wish to express gratitude to Mr Donald Burton and Mr Raj Patel for their support during the fieldwork. A special word of thanks is given to the Soil and Water Management Division of KU Leuven for its technical and administrative support during the laboratory work.

CONFLICT OF INTEREST

We can confirm that there is no conflict of interest for this manuscript.

AUTHOR CONTRIBUTIONS

Mulenga Kalumba: Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing-original draft; writing-review and editing. **Brecht Bamps:** Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing-review and editing. **Imasiku Nyambe:** Conceptualization; funding acquisition; methodology; project administration; resources; supervision; validation; writing-review and editing. **Stefaan Dondeyne:** Conceptualization; funding acquisition; methodology; project administration; resources; supervision; validation; writing-review and editing. **Jos Van Orshoven:** Conceptualization; funding acquisition; methodology; project administration; resources; supervision; validation; writing-review and editing.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Mulenga Kalumba https://orcid.org/0000-0002-6766-9919

Brecht Bamps https://orcid.org/0000-0003-0959-5082

REFERENCES

Araya, S. N., & Ghezzehei, T. A. (2019). Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. *Water Resources Research*, *55*, 5715–5737. https://doi.org/10.1029/2018WR024357

Batjes, N. H., Ribeiro, E., & Van Oostrum, A. (2019). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, *164*, 1–46. https://doi.org/10.17027/isric-wdcsoils.20190901

Botula, Y., Van Ranst, E., & Cornelis, W. M. (2014). Pedotransfer functions to predict water retention for soils of the humid tropics: A review. *Revista Brasileira de Ciência do Solo*, *38*(1), 679–698.

Breiman, L. E. O. (1996). Bagging predictors. *Machine Learning*, *24* (2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. E. O. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carsel, R. F., & Parrish, R. S. (1988). Developing joint probability distributions of soil water retention characteristics. *Water Resources Research*, *24*(5), 755–769.

Corréa, J. C. (1984). Características físico-h ídr icas dos solos latossolo amarelo, podzõlico vermelho-amarelo e podzol hidromórfico do estado do amazonas. *Pesquisa Agropecuaria Brasileira*, *19*(3), 347–360.

Cresswell, H. P., & Paydar, Z. (2000). Functional evaluation of methods for predicting the soil water characteristic. *Journal of Hydrology*, *227*(1–4), 160–172. https://doi.org/10.1016/S0022-1694(99)00178-X

Espino, A., Mallants, D., Vanclooster, M., & Feyen, J. (1996). Cautionary notes on the use of pedotransfer functions for estimating soil hydraulic properties. *Agricultural Water Management*, *29*(3), 235–253. https://doi.org/10.1016/0378-3774(95)01210-9

FAO. (2012). Crop Evapotranspiration (Guidelines for computing crop water requirement). *Irrigation and drainage*, *56*, 163

Gijsman, A. J., Jagtap, S. S., & Jones, J. W. (2003). Wading through a swamp of complete confusion: How to choose a method for estimating soil water retention parameters for crop models. *European Journal of Agronomy*, *18*, 77–106. https://doi.org/10.1016/S1161-0301(02)00098-9

Gupta, S. C., & Larson, W. E. (1979). Estimating soil-water retention characteristics from particle-size distribution, organic-matter percent, and bulk-density. *Water Resources Research*, *15* (6), 1633–1635.

Haghverdi, A., Cornelis, W. M., & Ghahraman, B. (2012). A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. *Journal of Hydrology*, *442–443*, 46–54. https://doi.org/10.1016/j.jhydrol.2012.03.036

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

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

IUSS Working Group WRB. (2015). *World Reference Base for Soil Resources 2014*, *update 2015*. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. FAO, Rome. https://doi.org/10.1017/S0014479706394902

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software*, *15*(9), 1–28.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer: New York, NY. https://doi.org/10.1007/978-1-4614-6849-3

Lamorski, K., Pachepsky, Y., Sławiński, C., & Walczak, R. T. (2008). Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. *Soil Science Society of America Journal*, *72*(5), 1243–1247. https://doi.org/10.2136/sssaj2007.0280N

Lantz, B. (2013). *Machine learning with R*. Birmingham: Packt Publishing Ltd. Retrieved from www.packtpub.com

Leenaars, J. G. B., van Oostrum, A. J. ., & Gonzalez, M. R. (2014). *Africa Soil Profiles Database*, *Version 1.2. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset)*. ISRIC Report 2014/01. Africa Soil Information Service (AfSIS) project and ISRIC - World Soil Inform. Wageningen, The Netherlands

Li, Y., Chen, D., White, R. E., Zhu, A., & Zhang, J. (2007). Estimating soil hydraulic properties of Fengqiu County soils in the North China plain using pedo-transfer functions. *Geoderma*, *138*(3–4), 261–271. https://doi.org/10.1016/j.geoderma.2006.11.018

Maclean, A. H., & Yager, T. U. (1970). Available water capacities of Zambian soils in relation to pressure plate measurements and particle size analysis. *Soil Science*, *113*(1), 23–29.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

McNeill, S. J., Lilburne, L. R., Carrick, S., Webb, T. H., & Cuthill, T. (2018). Pedotransfer functions for the soil water characteristics of New Zealand soils using S-map information. *Geoderma*, *326* (April), 96–110. https://doi.org/10.1016/j.geoderma.2018.04.011

Minasny, B., & Hartemink, A. E. (2011). Predicting soil properties in the tropics. *Earth Science Reviews*, *106*, 52–62. https://doi.org/10.1016/j.earscirev.2011.01.005

Minasny, B., Mcbratney, A. B., & Bristow, K. L. (1999). Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma*, *93*, 225–253.

Nemes, A., Schaap, M. G., & Wösten, J. H. M. (2003). Functional evaluation of pedotransfer functions derived from different scales of data collection. *Soil Science Society of America Journal*, *67*(4), 1093–1102. https://doi.org/10.2136/sssaj2003.1093

Nguyen, P. M., Haghverdi, A., De Pue, J., Botula, Y., Le, K. V., Waegeman, W., & Cornelis, W. M. (2017). Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. *Biosystems Engineering*, *153*(10), 12–27. https://doi.org/10.1016/j.biosystemseng.2016.10.013

Pham, K., Kim, D., Yoon, Y., & Choi, H. (2019). Analysis of neural network based pedotransfer function for predicting soil water characteristic curve. *Geoderma*, *351*(August 2018), 92–102. https://doi.org/10.1016/j.geoderma.2019.05.013

Pringle, M. J., Romano, N., Minasny, B., Chirico, G. B., & Lark, R. M. (2007). Spatial evaluation of pedotransfer functions using wavelet analysis. *Journal of Hydrology*, *333*(2–4), 182–198. https://doi.org/10.1016/j.jhydrol.2006.08.007

R Core Team. (2018). *R: A luanguage and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raes, D., Steduto, P., Hsiao, T. C., & Fereres, E. (2009). Aquacrop - the FAO crop model to simulate yield response to water: II. Main algorithms and software description. *Agronomy Journal*, *101*(3), 438–447. https://doi.org/10.2134/agronj2008.0140s

Rawls, W. J., Brakensiek, D. l., & Saxton, K. E. (1982). Estimation of soil water properties. *Transactions of ASAE*, *25*, 1316–1320.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining Knowl Discov*, *8*(1), 1–18. https://doi.org/10.1002/widm.1249

Saxton, K. E., & Rawls, W. J. (2006). Soil water characteristic estimates by texture and organic matter for hydrologic solutions.

Soil Science Society of America Journal, *70*, 1569–1578. https://doi.org/10.2136/sssaj2005.0117

Saxton, K. E., Rawls, W. J., Romberger, J. S., & Papendick, R. I. (1986). Estimating generalized soil-water characteristics from texture. *Soil Science Society of America Journal*, *50*(4), 1–18. https://doi.org/10.2136/sssaj1986.03615995005000040054x

Schaap, M. G., Leij, F. j., & van Genuchten, M. T. (1998). Neural network analysis for hierarchical prediction of soil hydraulic properties. *Soil Science Society of America Journal*, *62*(4), 847–855.

Schaap, M. G., Leij, F. j., & van Genuchten, M. T. (2001). ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, *251*, 163–176. Retrieved from www.cals.arizona.edu/research/rosetta/download/rosetta.pdf

Sequeira, C. H., Wills, S. A., Seybold, C. A., & West, L. T. (2014). Geoderma predicting soil bulk density for incomplete databases. *Geoderma*, *213*, 64–73. https://doi.org/10.1016/j.geoderma.2013.07.013

Sharma, A., Sandooja, B., & Yadav, D. (2013). Extreme machine learning: Feed forward networks. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(8), 1366–1371.

Steduto, P., Hsiao, T. C., Raes, D., & Fereres, E. (2009). Aquacrop - the FAO crop model to simulate yield response to water: I. concepts and underlying principles. *Agronomy Journal*, *101*(3), 426–437. https://doi.org/10.2134/agronj2008.0139s

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., & Rajkai, K. (2019). Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics. *Hydrology and Earth System Sciences*, *23*, 2615–2635. https://doi.org/10.5194/hess-23-2615-2019

Timlin, D. J., Pachepsky, Y. A., Acock, B., & Whisler, F. (1996). Indirect estimation of soil hydraulic properties to predict soybean yield using GLYCIM. *Agricultural Systems*, *52*(2), 331–353.

Tomasella, J., & Hodnett, M. G. (1996). Soil hydraulic properties and Van Genuchten parameters for an oxisol under pasture in central Amazonia. In J. H. C. Gash, C. A. Nobre, & R. L. Victoria, (Eds), *Amazonian Deforestation and Climate*. (pp. 101–124). New York: John Wiley and Sons.

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & Van Hijum, S. A. F. T. (2012). Data mining in the life sciences with random forest: A walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326. https://doi.org/10.1093/bib/bbs034

Twarakavi, N. K. C., Šimůnek, J., & Schaap, M. G. (2009). Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. *Soil Science Society of America Journal*, *73*(5), 1443–1452. https://doi.org/10.2136/sssaj2008.0021

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, *910*(11), 1–37. https://doi.org/10.3390/w11050910

Valyon, J., & Horváth, G. (2005). A robust LS-SVM regression. *World Academy of Science, Engineering and Technology*, *7*, 148–153.

van den Berg, M., Klamt, E., van Reeuwijk, L. P., & Sombroek, W. G. (1997). Pedotransfer functions for the estimation of moisture retention characteristics of ferralsols and related soils. *Geoderma*, *78*, 161–180.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils 1. *Soil Science Society of America Journal*, *44*, 892–898.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., … Vereecken, H. (2017). Pedotransfer functions in earth system science: Challenges and perspectives. *Reviews of Geophysics*, *55*, 1199–1256. https://doi.org/10.1002/2017RG000581

Van Wart, J., Grassini, P., Yang, H., Claessens, L., Jarvis, A., & Cassman, K. G. (2015). Creating long-term weather data from thin air for crop simulation modeling. *Agricultural and Forest Meteorology*, *209–210*, 49–58. https://doi.org/10.1016/j.agrformet.2015.02.020

Van Wart, J., van Bussel, L. G. J., Wolf, J., Licker, R., Grassini, P., Nelson, A., … Cassman, K. G. (2013). Use of agro-climatic zones to upscale simulated crop yield potential. *Field Crops Research*, *143*, 44–55. https://doi.org/10.1016/j.fcr.2012.11.023

Wassar, F., Gandolfi, C., Rienzner, M., Chiaradia, E. A., & Bernardoni, E. (2016). Predicted and measured soil retention curve parameters in Lombardy region north of Italy. *International Soil and Water Conservation Research*, *4*(3), 207–214. https://doi.org/10.1016/j.iswcr.2016.05.005

Wösten, J. H. M., Finke, P. A., & Jansen, M. J. W. (1995). Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. *Geoderma*, *66*(3–4), 227–237. https://doi.org/10.1016/0016-7061(94)00079-P

Wösten, J. H. M., Pachepsky, Y. A., & Rawls, W. J. (2001). Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology*, *251*(3–4), 123–150. https://doi.org/10.1016/S0022-1694(01)00464-4

Yi Lin, J., Cheng, C., & Wing Chau, K. (2007). Using support vector machines for long-term discharge prediction. *Hydrological Sciences*, *51*(4), 599–612.

Zhang, Y., & Schaap, M. G. (2017). Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3). *Journal of Hydrology*, *547*, 39–53. https://doi.org/10.1016/j.jhydrol.2017.01.004

How to cite this article: Kalumba M, Bamps B, Nyambe I, Dondeyne S, Van Orshoven J. Development and functional evaluation of pedotransfer functions for soil hydraulic properties for the Zambezi River Basin. *Eur J Soil Sci*. 2020; 1–16. https://doi.org/10.1111/ejss.13077