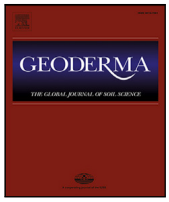
Geoderma 439 (2023) 116661

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



The appraisal of pedotransfer functions with legacy data; an example from southern Africa

C. Miti [a,b,*], V. Mbanyele [c], T. Mtangadura [c], N. Magwero [d], W. Namaona [d], K. Njira [d], I. Sandram [d], P.N. Lubinga [b], C.B. Chisanga [b], P.C. Nalivata [d], J.G. Chimungu [d], H. Nezomba [c], E. Phiri [b], R.M. Lark [a]

[a] School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK
[b] University of Zambia, School of Agricultural Sciences, Great East Road Campus, Lusaka, Zambia
[c] University of Zimbabwe, Department of Soil Science and Agricultural Engineering, Mount Pleasant, Harare, Zimbabwe
[d] Lilongwe University of Agriculture and Natural Resources, Crop and Soil Sciences Department, Bunda Campus, Lilongwe, Malawi

ARTICLE INFO

Keywords:
Pedotransfer functions
Field capacity
Wilting point
Soil legacy data
Scales of prediction
Linear mixed model

ABSTRACT

Predictions of soil hydraulic properties by pedotransfer functions (PTFs) must be treated with caution when they are used in an application domain which differs from the domain of their original development and calibration. However, in some settings, scientists may have little alternative but to use PTFs calibrated elsewhere. In this paper we consider how legacy data can be used to evaluate PTFs in new regions, paying particular attention to the challenges that arise when, as is often the case, the legacy data are not obtained by independent random sampling, and may be clustered at multiple scales. We undertook this work in southern Africa (Zimbabwe, Zambia and Malawi) where PTFs have been little-used, despite the scarcity of direct measurements of the soil properties of interest. We evaluated the extent to which existing PTFs provide a useful tool for the prediction of soil moisture content at field-capacity (−33 kPa) and permanent wilting-point (−1500 kPa) at different spatial scales. Soil legacy data for Zambia, Zimbabwe and Malawi were collated from various sources and PTFs from temperate and tropical domains were evaluated. We examined error variance components of predictions at within-profile, within-site and between-site scales; and estimated their mean errors. In general the better-performing PTFs (with respect to bias and the size of the error variance components) were ones calibrated with data from a tropical domain. This was most apparent at −1500 kPa. However, not all PTFs calibrated with data on tropical soils performed well, and predictions from some PTFs calibrated over a temperate domain were better at −33 kPa. The observations were spatially clustered, with data from different depth intervals in the same profile, from profiles in the same experimental site or farm, and from clusters across the region. This enabled us to show, with an appropriate mixed model analysis, that PTFs which effectively capture regional-scale variation may be less useful for predicting variation within a profile. We propose that such studies, based on legacy data, and with a suitable linear mixed model, should be used to screen PTFs of any provenance before their wider application.

1. Introduction

The application of predictive models in soil science has increased in recent years. These include models which simulate crop growth, soil erosion, catchment hydrology and effects of climate change (Wösten et al., 2013). However, these models require soil hydraulic properties and data on these are typically very scarce due to the costs of measurement, the specialist equipment required, the time required (e.g. to equilibrate a soil sample at the largest tensions) and the variability of soil (Minasny and Hartemink, 2011; Wösten et al., 2013). Pedotransfer functions (PTFs) are an alternative to routine measurement of such soil properties. PTFs are empirical predictive relationships between easy-to measure soil properties (e.g. soil texture, bulk density or organic matter), which are commonly recorded in soil surveys, and costly, time-consuming properties (e.g. the soil moisture characteristic curve) which are required for process models.

Various PTFs have been developed globally (Gupta and Larson, 1979; Rawls and Brakensiek, 1985; Aina and Periaswamy, 1985; Saxton et al., 1986; Dijkerman, 1988; Vereecken et al., 1989; Van den Berg et al., 1997; Schaap et al., 2001; Saxton and Rawls, 2006; Nemes

* Corresponding author at: School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK.
E-mail address: chawezi.miti@nottingham.ac.uk (C. Miti).

https://doi.org/10.1016/j.geoderma.2023.116661
Received 29 September 2022; Received in revised form 4 August 2023; Accepted 8 September 2023
Available online 19 September 2023

et al., 2008; Minasny and Hartemink, 2011; Botula et al., 2012). These provide an opportunity to deploy process models in regions where, hitherto, they have not been used because of the lack of soil hydraulic information. In particular there is unrealized potential in tropical regions to improve the efficiency of cropping systems, and their resilience to climate change, by developing strategies for soil water management which are site-specific, reflecting local weather conditions and soil properties. Agronomists, soil scientists, farmers and other land managers may also use soil hydraulic properties to make timely and reliable predictions of soil water dynamics, particularly in the context of high rainfall variability due to climate change, and so to support efficient use of irrigation water.

Most available PTFs have been calibrated on data sets such as HYPRES (Wösten et al., 1999) and WISE (Batjes, 1996). These are dominated by soil samples from temperate environments. Many authors have expressed caution about the transfer of PTFs developed with data on soils in one environment to a different setting (Givi et al., 2004; Botula et al., 2012; Shein and Arkhangel'Skaya, 2006; Tomasella and Hodnett, 2004). This is an important consideration in the tropics as few PTFs have been developed in tropical regions. Furthermore, those developed in tropical settings have typically been based on small data sets from restricted regions compared to the data sets available to develop PTFs for soils in temperate environments.

In regions where few or no PTFs have been developed from local data, legacy soil information can be used to evaluate the 'portability' of established PTFs into a new setting, and might also be used to develop locally-adapted parameter sets. However, by their nature, legacy data sets are not collected according to a single sample design, and they may reflect the spatial distribution of experimental farms or past survey campaigns. In such a case, the linear mixed model can be used both in the evaluation of PTFs and in the re-estimation of local versions from such data.

The objective of this study was to assess the applicability, across contrasting regions of Zambia, Zimbabwe and Malawi, of PTFs developed in various calibration domains, including temperate and tropical domains, the latter including both tropical Africa and the broader tropics. This was done with a range of validation data from legacy sources. These allowed us to assess the efficacy of the PTFs for predicting soil water content over different spatial scales from regional to within-profile.

2. Methods

2.1. Data collation and editing

Three teams, based in Zambia, Zimbabwe and Malawi, undertook a systematic search for legacy data on physical and compositional properties of soils:– particle size distribution, organic carbon content, dry bulk density and hydraulic data including gravimetric or volumetric water content (VWC) at specified tensions, or at field capacity or wilting point, defined either as −33 kPa or −1500 kPa respectively or, in some cases determined by an irrigation method (saturating the soil and allowing it to drain freely for 48 h) or the sunflower method (measuring the tension at which sunflower plants grown in the soil begin to wilt). Salter and Haworth (1961) outline these methods. The data sources comprised peer-reviewed publications, MSc and PhD theses, internal reports and the WOSIS database (Batjes et al., 2017) as it stood in 2016, from which records with a CC-BY Open Access licence, corresponding to sites in Zimbabwe, Zambia and Malawi were extracted. All sources are recorded in a separate reference list provided as Appendix A in the supplementary material.

Each observed object in the final data collection was a set of data values for a single soil sample, collected from a particular depth interval or horizon with recorded upper and lower depths. Each object was given a unique *entry* code. In most cases soil data were available for multiple horizon or depth intervals from a single sample point. All objects from a single such set were given a unique *profile* code. Many soil profiles in the data set had longitude and latitude recorded. For those that did not, location information (e.g. from a local large-scale map, descriptive notes or basic information on the farm or research station from which the soil samples had been collected) was used to obtain location coordinates using all information provided on the sample location in the original source material. For example, one profile was described as occurring '43 km NE of Livingstone, Southern Province along the road to Lusaka' (Zambia).

These coordinates were not treated as unique for each sample, but comprise the best possible post-hoc identification of the central *location* of a cluster of samples from a named study site, research station or farm. The coordinates were required so that the clustered distribution of sample sites could be appropriately reflected in the data analysis by a between-location random effect. Spatial clustering of samples is to be expected in such a legacy data set where measurements have been made at fields in particular research stations, farms selected for experimental studies, or sites where a detailed soil survey has been made.

In addition to an entry code, profile code and location code and coordinates, care was taken to identify all data used to obtain any of the PTFs which had been identified for evaluation. Any such datum was given an appropriate PTF code so that it could be excluded from evaluation of the specific PTF.

In collating the data, particular attention was paid to the following possible sources of error. First, the definition of particle size classes was carefully checked in the data source. It is well-known that the silt/sand limit differs between the conventions of USDA or FAO and others such as the ISSS (Landon, 2014). Some of the sources presented particle size data according to more than one convention. Many of the sources provided more precise particle size categories (e.g. coarse, medium and fine sand), according to specific conventions. Several particle size variables were therefore recorded separately in the database, corresponding to the different size categories of different conventions. Second, all units were checked and standardized (e.g. for soil organic carbon which was recorded for all observations in units of percent by mass). Care was taken as to whether the source reported soil organic carbon or soil organic matter. The reported variable was recorded for any sample. In the use of the data if an interconversion was needed between these variables it was done on the assumption that SOM = SOC × 1.72 (Landon, 2014), unless the source recorded that the conversion had already been made with a different factor. Third, the basis for all measurements was carefully checked. In many cases, for example, gravimetric water content was reported alongside a dry bulk density value by which it could be converted to volumetric water content. Because of these criteria, by no means all data found in the initial screening were used in the analyses reported below. We checked that all measurements of water content of the soil at field capacity/−33 kPa were made on intact cores, as water retained at this tension is partly in pore space between soil aggregates.

A total of 602 horizon entries were extracted from 129 profiles at 90 locations. The locations are plotted over the borders of Zimbabwe, Zambia and Malawi in Figure S1(a) of the Supplementary Material. The Soil Reference Group (SRG) from the World Reference Base classification (IUSS Working Group WRB, 2006) was extracted at each location from the Soil Atlas of Africa shapefiles (Jones et al., 2013). The percentages of locations in each of 13 SRG are shown in Figure S1(b) in the Supplementary Material. Note that about 30% of the soils are Luvisols (the largest group) followed by Arenosols at just over 15% and Acrisols and Ferralsols at about 10% each. The range of longitudes and latitudes for the locations are {22.13, 34.01} degrees and {−21.83, −8.83} degrees respectively. Climate information from the locations was extracted from the AfroGRID data set, Schon and Koren (2022). The mean annual temperature over all the sites was 21.5 degrees C, with a range of {18.8, 25.2} and the mean annual precipitation was 846 mm with a range of {430, 1253}.

Table 1
Summary of pedotransfer functions used in this study.

Author	Broad domain	Comment on domain	Predictors for θ_{-33}					Predictors for θ_{-1500}				
			Sand	Silt	Clay	SOM	BD	Sand	Silt	Clay	SOM	BD
Botula et al. (2013)	Tropical	Central Africa-Lower Congo	•					•		•		•
Dijkerman (1988)	Tropical	West Africa-Sierra Leone	•							•		
Lal (1978)	Tropical	Nigeria			•					•		
MacLean, Yager (1972)	Tropical	Southern Africa-Zambia	•[c]	•	•	•[b]			•	•	•[b]	
Minasny, Hartemink (2011)	Tropical	Tropical soils	•				•			•[e]	•	
Miti (2017)	Tropical	Southern Africa-Zambia		•	•	•				•	•	
Oliveira et al. (2002)	Tropical	S.America- N.E. Brazil		•	•			•	•	•		•
Pidgeon (1972)	Tropical	East Africa -Uganda		•	•	•[b]			•	•	•[b]	
Rawls and Brakensiek (1982)[d]	Temperate	North America-USA	•		•	•				•	•	
			•			•						
Saxton and Rawls (2006)	Temperate	North America-USA	•[a]		•[a]	•[a]		•[a]		•[a]	•[a]	
van den Berg et al. (1997)	Tropical	Tropical Oxisols							•[a]	•[a]		•[a]

[a] The predictors are in a non-linear combination.
[b] SOC is the predictor.
[c] Coarse sand (FAO definition).
[d] Second PTF for θ_{-33} is denoted Rawls et al (1982) b.
[e] Clay and Clay^2.

2.2. Initial exploratory analysis

The literature on PTFs was reviewed to identify PTFs developed with soil observations within the region, and others used in studies within it. We identified those PTFs for which some subset of our data included all the predictor variables. Some were excluded because the published PTFs did not produce predictions for VWC at the tensions for which we had validation data, for example, the PTFs of Mugabe (2004). The PTFs evaluated in this study are presented in Table 1. In Figure S2 we show the range of values for the key predictor variables in the calibration data set for each of the published PTFs, and also show the range and quartiles of these values for our legacy data.

For any specified PTF, the appraisal using the assembled database started with exploratory analysis. First we extracted the subset of observations in the database for which all the required predictor variables were available. We also removed any data which had been used to develop the particular PTF. We then used the PTF to make predictions of volumetric water content at the tension, h, for which it is specified. We then made a plot of the observed volumetric water content at the specified tension, θ_h against the value predicted by the PTF, $\breve{\theta}_h$. The bisector (1:1 line) was drawn on the plot. Inspection of the plots shows obvious biases, unphysical predictions and any shrinkage effects (regression to the mean), with small VWC over-predicted and large VWC under-predicted.

The PTF error for each observation was then computed as

$$\varepsilon = \theta_h - \breve{\theta}_h, \tag{1}$$

so a positive error means that the observed water content exceeds the predicted value. The histograms of the errors and their summary statistics were examined. In order better to understand possible sources of error the prediction error was plotted against the mid-depth of the sampled horizon or interval, the sand, silt and clay content, the soil organic carbon content and the bulk density.

Finally, we examined the source publication for each PTF and, where possible, identified the range of values for each predictor variable in the set which had been used to estimate the PTF parameters. These ranges were plotted along with the range of values for the same variables in our database, and a box plot (showing the median values and first and third quartiles).

2.3. Statistical modelling

We first present a general statistical model for these data, before explaining how it was used for evaluation and for re-parameterization.

Consider an observation of some variable, Z, recorded for horizon i in profile j at location k in the data set, where location k has coordinates $\mathbf{s}_k$. We may specify a linear mixed model (LMM) for this variable of the form

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\psi}\left(\mathbf{s}_k\right) + \boldsymbol{\omega}_j + \boldsymbol{\varepsilon}_i, \tag{2}$$

where the first term on the right-hand side is an expression of the mean value of Z in terms of a fixed effect variable or variables in the design matrix $\mathbf{X}$, and a fixed effect coefficient or coefficients in the vector $\boldsymbol{\beta}$. In the simplest case where the mean is a constant the fixed effect is a vector of ones, and the coefficient is the mean value. In a more complex case, where Z is a soil property for which we are re-parameterizing a PTF, the fixed effects would constitute the easy-to-measure soil properties that are predictors in the PTF, and the fixed effect coefficients would be the re-estimated PTF coefficients.

The remaining terms on the right-hand side of Eq (2) are random effects, all assumed to be of mean zero and with a normal distribution. The first term is a spatially correlated random variable which takes fixed values over all observations within a given location. It is assumed to be second-order stationary in that the covariance of any two values $\boldsymbol{\psi}\left(\mathbf{s}_1\right)$ and $\boldsymbol{\psi}\left(\mathbf{s}_k\right)$ exists and can be expressed as a function of the lag vector $\mathbf{s}_1 - \mathbf{s}_2$ which denotes the difference between their locations, independent of these locations themselves. In this study, which encompasses widely distributed locations in both longitude and latitude, we could not project all locations onto a rectilinear grid, and so we used a covariance function of great-circle distance between locations approximated on the sphere. These distances were obtained using the distVincentySphere function from the geosphere package for the R platform (Hijmans et al., 2017). Given the relative sparsity of locations, and their irregular spatial distribution it was assumed that the covariance function was isotropic, i.e. its argument was just the spherical great circle distance, which we denote by $\left|\mathbf{s}_1 - \mathbf{s}_2\right|^{\odot}$. Specifically we assumed a Matérn covariance function (Stein, 1999) which takes the form:

$$C_\psi\left(\left|\mathbf{s}_1 - \mathbf{s}_2\right|^{\odot}\right) = \sigma_\psi^2 \left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1} \left(\frac{\left|\mathbf{s}_1 - \mathbf{s}_2\right|^{\odot}}{\phi}\right)^{\kappa} K_\kappa\left(\frac{\left|\mathbf{s}_1 - \mathbf{s}_2\right|^{\odot}}{\phi}\right), \tag{3}$$

where σ_ψ^2 is the between-location variance, $\Gamma(\cdot)$ is the Gamma function, $K_\kappa(\cdot)$ is a modified Bessel function of the second kind of order κ, and the parameters κ and ψ are, respectively a dimensionless quantity that characterizes the smoothness of the spatial variation of $\boldsymbol{\psi}(\mathbf{s})$ and a spatial parameters with units of distance which, conditional on the value of the smoothness parameter, characterizes the distance over

which the variable is spatially correlated. Gneiting (2013) shows that the Matérn covariance function is positive definite, i.e. valid, for real positive lag distances on the sphere provided that $0 < \kappa \leq 0.5$.

The term $\boldsymbol{\omega}_j$ in Eq [(2)] is a normal random variable of mean zero, independently and identically distributed with variance σ^2_ω. It is a between-profile within-location random effect, so takes a constant value over all observations in profile j. Similarly, the final term $\boldsymbol{\epsilon}_i$ is a normal random variable of mean zero, independently and identically distributed with variance σ^2_ε. It is a within-profile residual term, and so will include the effects of any analytical error.

Because of the nested structure of this model (observations within profiles within sites), and on the assumption that the three random effects are mutually independent, one may write a covariance matrix for a set of n observations, $\mathbf{z} = \{z_1, z_2, \ldots, z_n\}^{\mathrm{T}}$. These observations have mean $\mathbf{X}\boldsymbol{\beta}$ where the first term is a $n \times p$ design matrix with a set of p values of the fixed effects the second term is a $p \times 1$ matrix of fixed effects coefficients. One may specify a component of the overall $n \times n$ covariance matrix of the random effects attributable to the spatially correlated random variable $\boldsymbol{\psi}(\mathbf{s})$. This matrix is $\mathbf{C}_\psi$ where

$$\mathbf{C}_\psi[l, m] = C_\psi\left(\left|\mathbf{s}_l - \mathbf{s}_m\right|^{\odot}\right). \tag{4}$$

The vectors $\mathbf{s}_l$ and $\mathbf{s}_m$ denote the locations of the lth and mth observation (which might be identical, for example for observations in two horizons of the same profile or two profiles at the same location, and the covariance function on the right side is defined in Eq [(3)] given values of the parameters σ^2_ψ, κ and ϕ.

If there are n_P locations then one may specify a $n \times n_P$ design matrix for the random effect $\boldsymbol{\omega}_j$. This matrix, $\mathbf{H}_\omega$, has value 1 in the kth column of the mth row if the mth observation is in the kth profile, all other entries in that row are zero. The covariance matrix for the random effect $\boldsymbol{\omega}_j$ can then be written as:

$$\mathbf{C}_\omega = \sigma^2_\omega \mathbf{H}_\omega \mathbf{H}^{\mathrm{T}}_\omega. \tag{5}$$

Finally, given the variance of the random effect $\boldsymbol{\epsilon}_i$, which we denote by σ^2_ε, the covariance matrix for this residual term is given by

$$\mathbf{C}_\varepsilon = \sigma^2_\omega \mathbf{1}_n, \tag{6}$$

where $\mathbf{1}_n$ is the $n \times n$ identity matrix.

This notation now allows us to specify the overall covariance matrix for $\mathbf{Z}$ as

$$\mathbf{C} = \mathbf{C}_\psi + \mathbf{C}_\omega + \mathbf{C}_\varepsilon, \tag{7}$$

assuming the mutual independence of the three random effects, and given values of the parameters $\sigma^2_\varepsilon 1$, σ^2_ω, σ^2_ψ, κ and ϕ.

These five variance parameters, so-called because they are the full set required to characterize the random effects in the model, can be estimated for any set of observation of the specified dependent variable, Z, and set of fixed effects variables in the design matrix $\mathbf{X}$. The preferred method to obtain these estimates is by residual maximum likelihood (Patterson and Thompson, 1971) which is straightforward for unbalanced data, and which reduces the bias due to the fact that the fixed effects coefficients in $\boldsymbol{\beta}$ are unknown. We do not provide a detailed account of REML in this paper, but refer the reader to other sources e.g. Lark et al. (2006),Verbeke (1997). When the REML estimates of the variance parameters have been obtained they can be used to compute an estimate of the covariance matrix $\mathbf{C}$, by the application of Eqs (3)–(7) above, which we denote by $\mathbf{C}^*$. Having obtained this we can use it to obtain estimates of the fixed effects coefficients by generalized least squares:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{C}^{*-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{C}^{*-1}\mathbf{z} \tag{8}$$

and the covariance matrix for the estimation error of these coefficients can be obtained by:

$$\mathbf{B} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{C}^{*-1}\mathbf{X}\right)^{-1} \tag{9}$$

In this study we obtained estimates of the variance parameters that minimize the negative log residual likelihood (REML estimate) using the optim function in base R (R Core Team, 2020), and specifically we used the L-BFGS-B (Low-memory Broyden–Fletcher–Goldfarb–Shanno, bounded algorithm) method (Byrd et al., 1995).

Rather than estimating the smoothness parameter κ by allowing it to vary freely with the other variance parameters, we followed Diggle and Ribeiro (2007) by using a profile method by which we fixed κ at a series of values and found the REML estimates of all other variance parameters, conditional on this fixed value, then selected the solution for which the residual negative log-likelihood was smallest. We considered values of $\kappa = 0.1, 0.2, 0.3, 0.4, 0.5$. By fixing $\kappa \leq 0.5$, we ensured that the covariance model was positive definite for points on the sphere (Gneiting, 2013).

2.4. Specific models

The modelling procedure described in the previous section was used to fit the following models in turn.

In the first model, M_0, the dependent variable, Z, was the measured value of volumetric water content at tensions of −33 kPa or −1500 kPa, θ_{-33} and θ_{-1500}, with a constant mean as the only fixed effect.

The second model, M_1, was for the prediction errors of the PTF, as defined in Eq (1) above. The only fixed effect specified was a constant mean error. This fixed effect was estimated by generalized least squares given the REML estimates of the random effects parameters, using Eq. (8). This estimate of the mean error, with a confidence interval, could be examined for evidence that predictions made by the PTF are biased overall. The variance components for each of the random effects were estimated, and then compared with the corresponding components for model M_0.

None of the PTFs were for a specified depth (unlike some in the literature, e.g. Hall et al. 1977), so we considered the possibility that there is a different mean error for topsoil and subsoil. For purposes of this analysis any horizon entirely in the 0–200 mm depth interval was regarded as topsoil, and any horizon entirely below 200 mm depth was regarded as subsoil. This is a somewhat arbitrary division selected because 200 mm was the 80th percentile of the lower bound of all horizons with the upper bound at 0 mm. A horizon spanning 200 mm depth was regarded as topsoil if half or more of its thickness was shallower than 200 mm, and subsoil otherwise. Under this alternative model, M_2, we computed mean errors for topsoil and subsoil predictions, and also noted whether the confidence interval for the difference between mean errors in the two sets of soils included zero.

The next analysis was a re-parameterization in which the PTF was refitted to the available data, M_3. The variance components for the random effects were compared with the corresponding error variance components. If the coefficients of the PTF are improved by re-fitting then we would expect the variance components for the re-parameterized model to be smaller than the error model components (overall), although this might not apply to all the components if there is scale-dependence. If re-parameterizing has little effect, or only changes the constant, then the variances for the re-parameterization and the original error model should all be very similar.

3. Results

Results are presented on the errors of the predictions of VWC at θ_{-33} and θ_{-1500} in turn. To assess the evidence for the performance of a particular PTF we examined the scatter plot of predicted and measured values (Fig. 1 and Figure S3 for θ_{-33}), we also considered the mean prediction error and its 95% confidence interval, and the sum of components of the variance of the error on a plot (Fig. 2 for θ_{-33}). This gives an indication of the bias and scatter in the PTF predictions. The plots of the variance components by scale (within-profile, within-site, between-site) then show the scale-dependence of

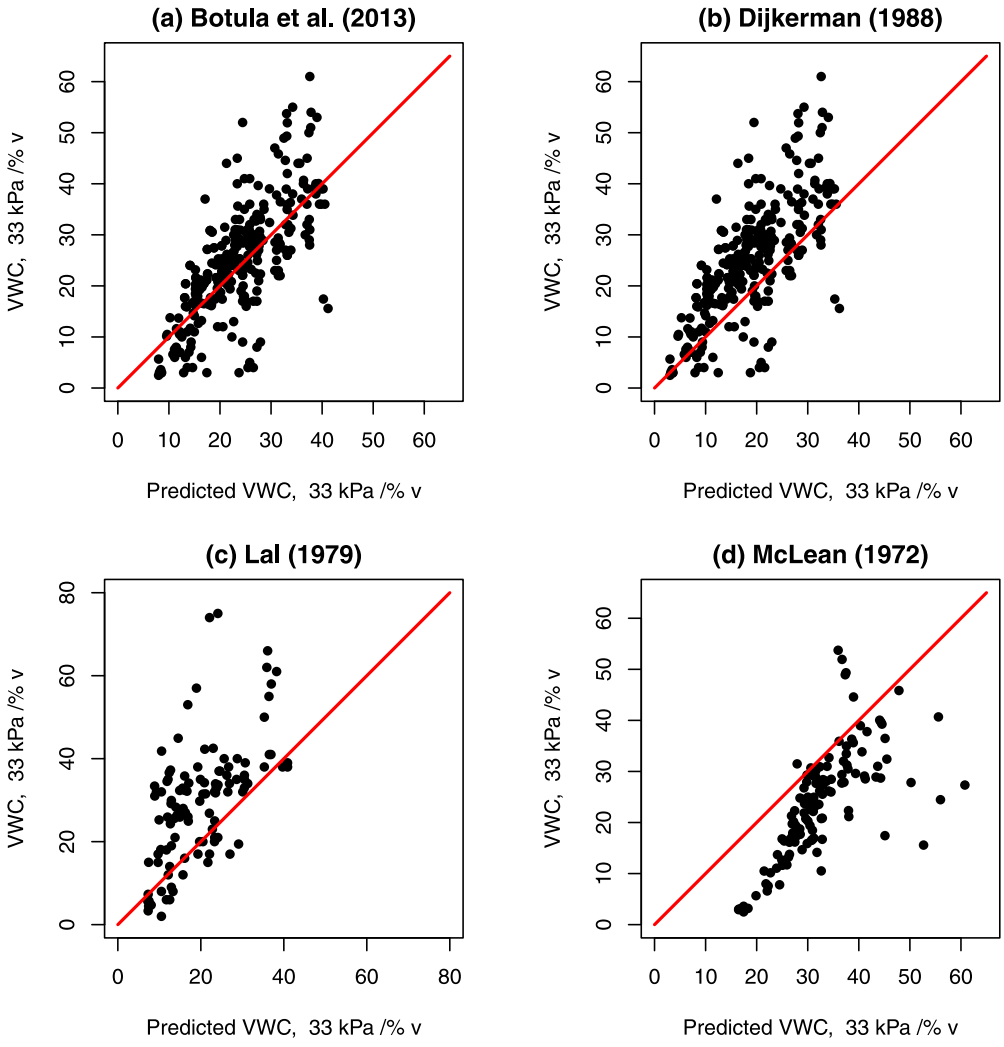


Fig. 1. Scatter plots of observed θ_{-33} (percent by volume) against the predictions by all the PTFs published for this variable by (a) Boutula et al. (2013); (b) Dijkerman (1988); (c) Lal (1979); (d) MacLean and Yager (1972). Note that the red line is the bisector (1:1 line) where the predicted and observed water contents are the same. Figure S2 in Appendix C (Supplementary Materials) shows the full set of these plots for all PTFs considered.

PTF uncertainty (Fig. 3 for θ_{-33}). A plot of the variance components for the error on refitting of each PTF to the data (estimation of the fixed effects coefficients) shown as a proportion of the variance component of the original VWC error, shows whether a substantial improvement is achieved by refitting the PTF (Fig. 4 for θ_{-33}). The findings from these four sets of results were then elucidated by examining the range of predictor values from the data used to fit the PTF, and a comparison of this with the range of values in the validation data (Figure S2), and a plot of PTF error against soil properties (e.g. Fig. 5).

In the discussion section we then examine overall PTF performance, trends and general findings.

3.1. PTF predictions of θ_{-33}

Fig. 1 shows the scatter plots of predicted VWC at −33 kPa (θ_{-33}) against the observed value for each of the PTFs considered. In Fig. 2 the mean error for the predictions for each PTF (θ_{-33}), with its 95% confidence interval, is plotted against the sum of components of the variance of the error. The values of these separate components are plotted in Fig. 3.

The PTF due to Botula et al. (2013) has the smallest absolute mean prediction error, which is not significantly different from zero. The mean error is slightly positive (i.e. a tendency to under-predict the VWC, Fig. 2.). The sum of error variance components for this PTF is relatively small (the third-equal smallest in the set, Fig. 2). Examination of the plot of measured θ_{-33} against the predicted value (Fig. 1a) shows a linear relationship, although with a tendency to 'shrinkage' with the variance of the predicted values smaller than that of the observations. The error variances are the same for the PTF due to Dijkerman (1988) (Fig. 2). These PTF both use sand content as the sole predictor, and have the same regression coefficient, but different intercepts. As a result the mean error for the predictions by the PTF of Dijkerman (1988) is significantly larger than zero.

The PTF due to Lal (1978) also has a significantly positive mean prediction error, and the sum of error variances is only slightly larger than for the PTFs of Botula et al. (2013) and Dijkerman (1988) (Fig. 2). This PTF uses only the clay content of the soil for prediction of θ_{-33}. For all three of these PTFs, the variance components of the prediction error are notably smaller than the corresponding variance component for the variable itself (Fig. 3). On reparameterization of these PTFs with the available data, the variance components of the random effects are only slightly smaller than the variance components of the error of the original PTF (Fig. 4). This reflects the fact that the re-estimated regression coefficients are very similar to those of the original PTF, although the difference in the intercepts is larger. For all three of these PTFs the difference between the mean errors in the topsoil and subsoil is significant, with θ_{-33} more under-predicted in the topsoil (Figure S4).

The range of values of sand content in the data set originally used by Botula et al. (2013) to fit the PTF is very close to that of the data

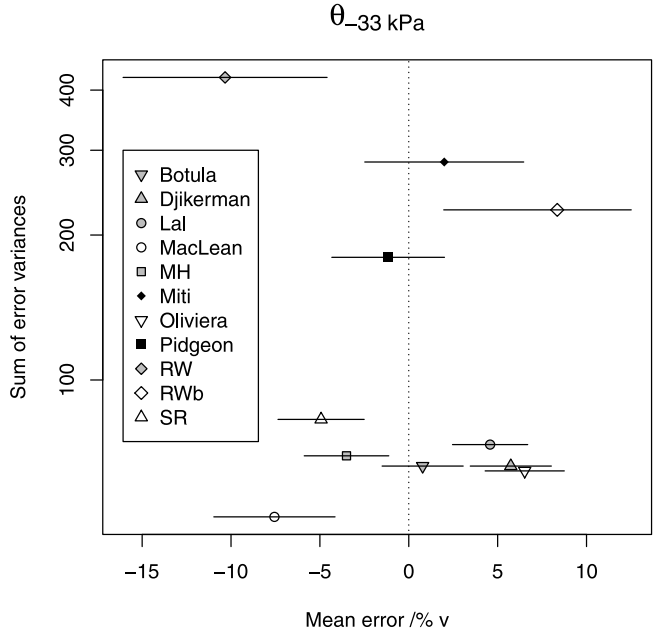


Fig. 2. For each PTF for θ_{-33}, the estimate of mean error, with error bars (twice the standard error), plotted against the sum of the variance components for the error. The PTF authors in the key are as follows. Botula (Botula et al., 2012); Djikerman (Dijkerman, 1988); Lal (Lal, 1978); MacLean (MacLean and Yager, 1972); MH (Minasny and Hartemink, 2011); Miti (Miti, 2017); Oliviera (Oliveira et al., 2002); Pidgeon (Pidgeon, 1972); RW (Rawls and Brakensiek, 1982); RWb (Rawls and Brakensiek, 1982); SR (Saxton and Rawls, 2006).

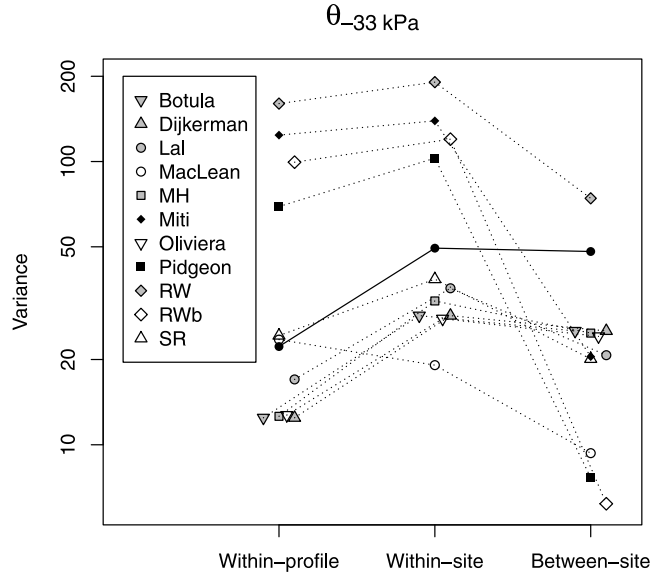


Fig. 3. Variances of the random effects for PTF error (θ_{-33}), from the model with a constant mean as the fixed effects. The solid black discs joined by an unbroken line show the corresponding variance components for θ_{-33}. The PTF authors in the key are as follows. Botula (Botula et al., 2012); Djikerman (Dijkerman, 1988); Lal (Lal, 1978); MacLean (MacLean and Yager, 1972); MH (Minasny and Hartemink, 2011); Miti (Miti, 2017); Oliviera (Oliveira et al., 2002); Pidgeon (Pidgeon, 1972); RW (Rawls and Brakensiek, 1982); RWb (Rawls and Brakensiek, 1982); SR (Saxton and Rawls, 2006).

used here for validation (Figure S2). However, the maximum sand content in the soils observed by Dijkerman (1988) was less than the third quartile of the values in the data used here (Figure S2). Despite this the regression coefficients for the two PTFs are identical although, as noted, the intercepts differ.

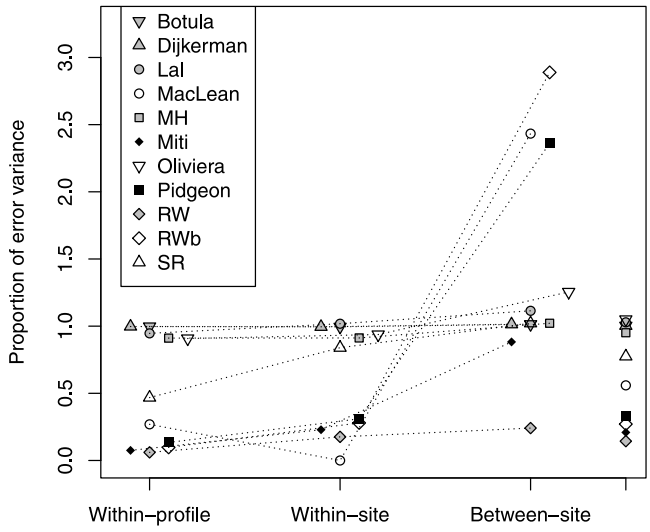


Fig. 4. Variances of the random effects from reparameterization of each θ_{-33}PTF expressed as a proportion of the corresponding variance components for that PTF's error. The PTF authors in the key are as follows. Botula (Botula et al., 2012); Djikerman (Dijkerman, 1988); Lal (Lal, 1978); MacLean (MacLean and Yager, 1972); MH (Minasny and Hartemink, 2011); Miti (Miti, 2017); Oliviera (Oliveira et al., 2002); Pidgeon (Pidgeon, 1972); RW (Rawls and Brakensiek, 1982); RWb (Rawls and Brakensiek, 1982); SR (Saxton and Rawls, 2006).

The PTF for θ_{-33} due to MacLean and Yager (1972) uses coarse sand, silt, clay and organic matter content of the soil for prediction. The sum of error variance components is smaller for this PTF than for any other but it does have a large negative and significant bias, over-predicting θ_{-33} by 7.6% v/v on average. Examination of the exploratory plots for this PTF, presented in Fig. 5, shows that the error is markedly related to the sand and clay content, with the largest over-prediction for the lighter-textured soils. This is also reflected in the difference between the mean errors in topsoil and subsoil (Figure S4), with a markedly larger mean over-prediction in the, generally lighter-textured, topsoil. This could be because the maximum sand content in the data set used by MacLean and Yager (1972) was smaller than for any of the other PTFs (Figure S2), and closer to the median value in our data than two their third quartile. As seen in Fig. 4, the variances of the within-profile and within-site random effects on reparameterization of the PTF of MacLean and Yager (1972) are small relative to the corresponding error variances, although the variance at the between-site scale is larger. This shows that there is some benefit in re-estimating the coefficients of this particular PTF, although less for capturing broad trends in the value of θ_{-33} than for representing variation with depth, or short-range variation within sites.

The two PTFs of Rawls and Brakensiek (1982) and Saxton and Rawls (2006) were both estimated with data from the temperate region (USA). The PTF of Rawls and Brakensiek (1982) denoted RW in the Figure legends, uses sand, clay and organic matter content as predictors. It has the largest sum of error variance components of all the PTFs for this variable (Fig. 2), and the largest single variance components for each random effect, all larger than the corresponding variance component for the variable itself (Fig. 3). It has the largest mean prediction error, which is negative, indicating a tendency to over-prediction (Fig. 2). Inspection of Figure S3(i), shows that many of the data are quite tightly clustered around the bisector. However, there are some very large predicted values, including several unphysical ones where the predicted VWC exceeds 100% by volume. The exploratory plots of PTF error for this PTF (Figure S5), show that this bias can be attributed to the effect of soil organic matter in the equation. Bias is introduced for soils with SOC in excess of around 5% by mass. This also explains the observation (Figure S4) that the mean over-prediction in topsoils is larger for this

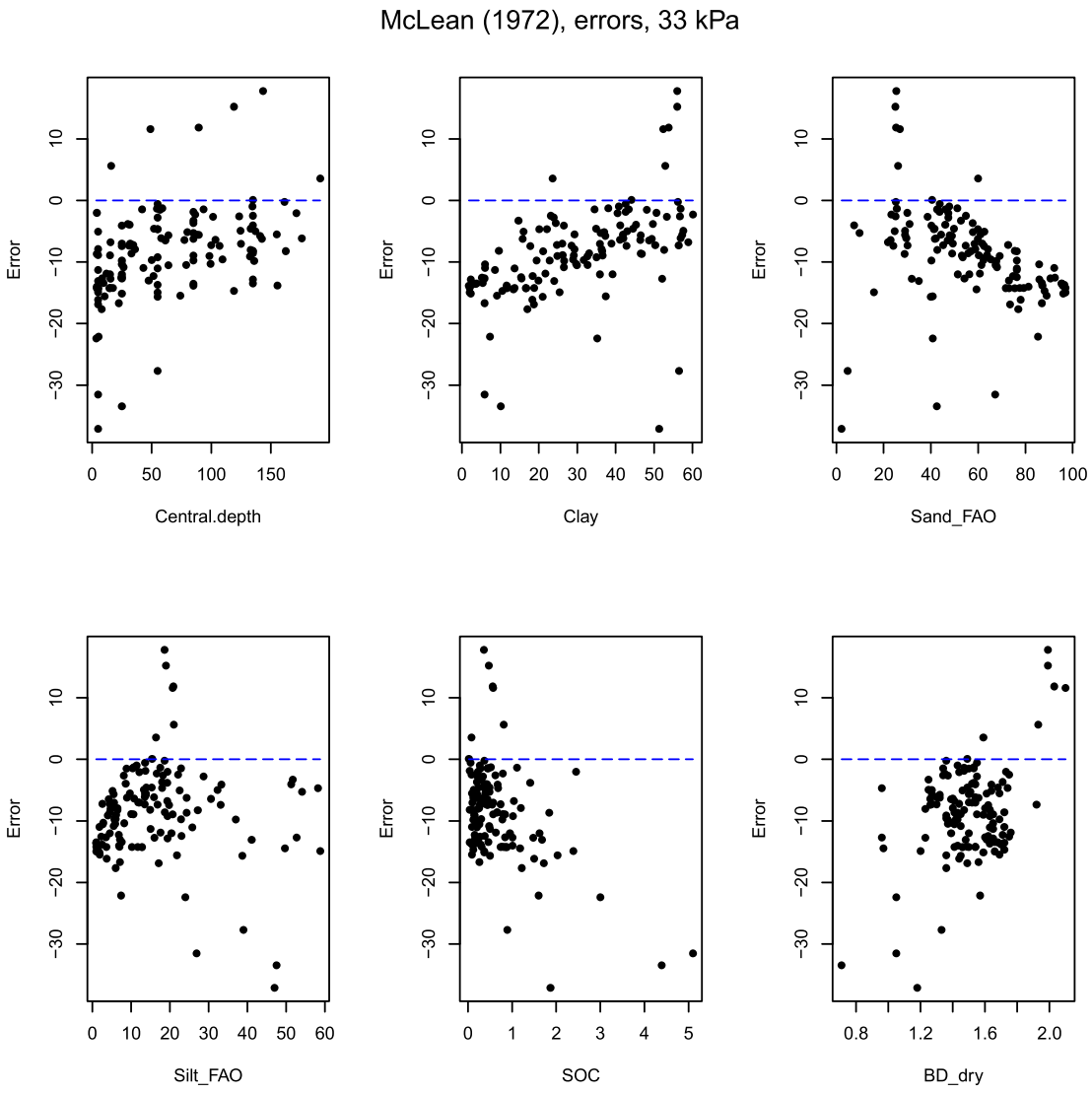


Fig. 5. Scatter plots of PTF error against the central depth of the observation, clay content, sand content, silt content, SOC content and dry bulk density. This is for the PTF for $\theta_{-33}$of MacLean and Yager (1972).

PTF than for any other, and significantly larger than the over-prediction in subsoils. The range of soil organic carbon contents in the soils used by Rawls and Brakensiek (1982) is smaller than in our data (Figure S2), but the evidence of bias in the PTF is seen at organic matter contents within the range of the data used by Rawls and Brakensiek (1982). It is notable that the variance components of the random effects in the reparameterization of this PTF are all markedly smaller than the corresponding variance components for PTF error (Fig. 4), showing that the model is poorly parameterized for the southern African setting. In particular the coefficient for soil organic matter in the original PTF, 0.0299, is an order of magnitude larger than in the re-fitted model (0.0017).

The PTF of Rawls and Brakensiek (1982) denoted RWb in the Figures, uses just sand and organic matter content for prediction. It has the largest positive bias of all the PTFs examined for this variable (under-prediction). Figure S3(j) shows that, while this PTF also produces over-predictions of the VWC, including unphysically large ones, the dominant effect is a positive bias which increases with the predicted value. Inspection of the exploratory plot for the PTF (Figure S6), shows that, in addition to the negative bias introduced for soils with relatively large organic matter content, there is a pronounced relationship between error and sand content, with a tendency to under-predict the water content of soils with a smaller sand content. Note in Figure S4 that the mean error for subsoil predictions is significantly larger than zero (under-prediction), and significantly different from the mean error for topsoil which is slightly negative. As with the first predictor for this variable from these authors, the sum of variance components for the random effects on reparameterization is smaller than the sum of variance components for the error (Fig. 4), although this does not hold for the between-site random effect considered alone.

The PTFs of Miti (2017) and of Pidgeon (1972), which both use silt, clay and soil organic carbon/matter as predictors, resemble the PTFs of Rawls and Brakensiek (1982) in that there are unphysical predictions associated with soils with larger soil organic content. Note that the soils used by Miti (2017) had a narrower range of organic matter contents than those used by Rawls and Brakensiek (1982). In both the mean error is not significantly different from zero, but there is a tendency for under-prediction of the VWC of soils with larger water content which do not have large organic content. The mean errors for predictions with these PTFs in the topsoil and subsoil are significantly different, with more negative values (over-prediction) for the topsoil (Figure S4). In both the relationship between error and sand content is similar, although less pronounced, to the second PTF of Rawls and Brakensiek (1982). The variance components of the within-profile and within-site random effects of the error are both very large (Fig. 3), but that for the between-site effect is smaller than for the variable itself. It is possible

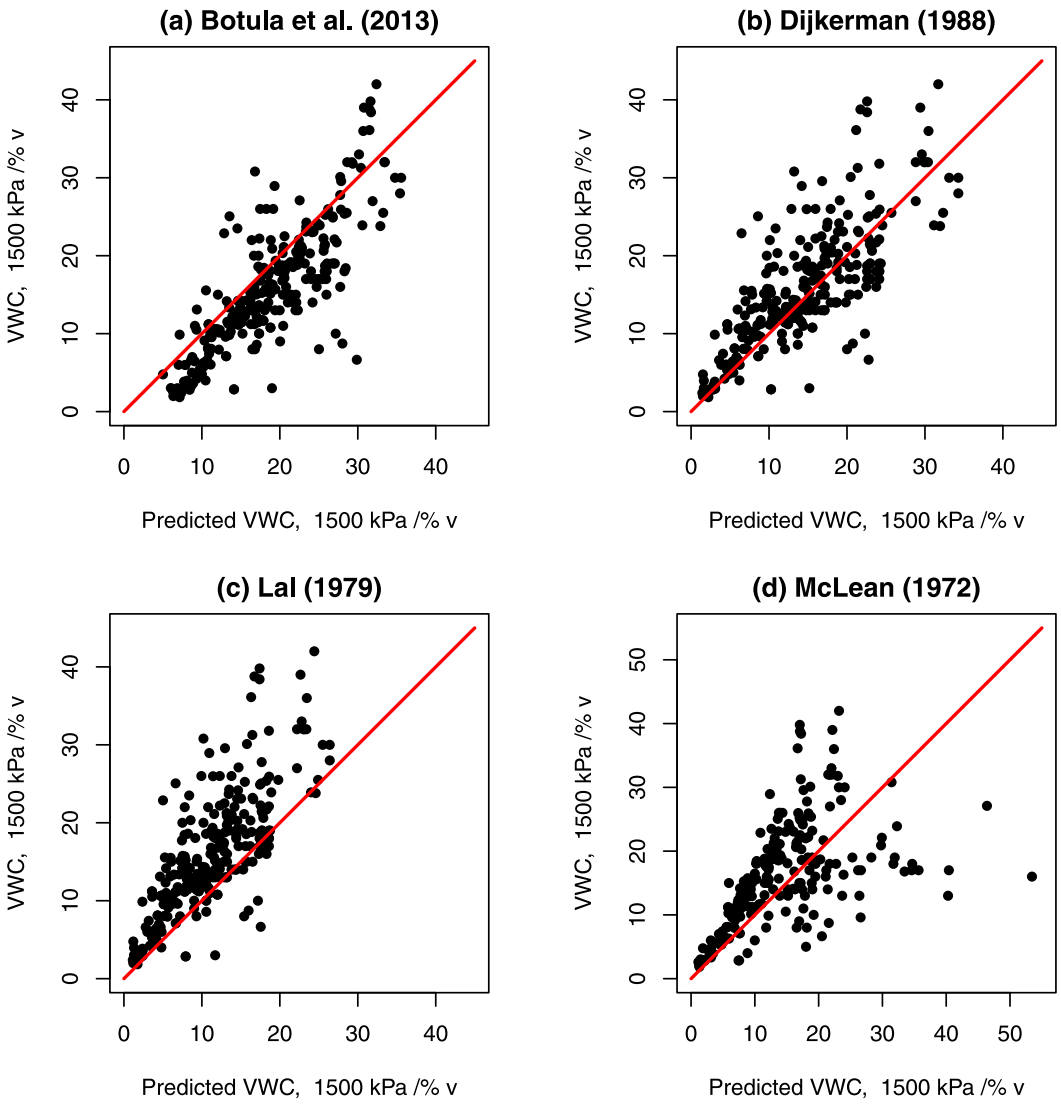


Fig. 6. Scatter plots of observed θ_{-1500} (percent by volume) against the predictions by all the PTFs published for this variable by (a) Botula et al. (2013); (b) Dijkerman (1988); (c) Lal (1979); (d) McLean (1972). Note that the red line is the bisector (1:1 line) where the predicted and observed water contents are the same. Figure S4 in Appendix C (Supplementary Materials) shows the full set of these plots for all PTFs considered.

that these larger error variances at the two shorter-scales, for the PTFs of Miti (2017), Pidgeon (1972) and Rawls and Brakensiek (1982) reflect errors introduced by variation in the organic carbon content of soil horizons of different depths, and fields within sites with contrasting management.

The PTF of Minasny and Hartemink (2011), fitted to data from a range of tropical soils, is the only one for θ_{-33} which uses bulk density as a predictor. The mean error is negative, and significantly different from zero, and there is no significant difference between the mean errors in the topsoil and subsoil. The error variance components are all smaller than the corresponding variance components for the variable itself, and the variances for the random effects in the reparameterization of the model are all only slightly smaller than the corresponding error variance components. The scatterplot in Figure S3(e) shows shrinkage, with a tendency for over-prediction for the soils with smaller than average observed θ_{-33}, and under-prediction for those with θ_{-33} above average.

The PTF due to Saxton and Rawls (2006) has a significant negative bias (Fig. 2), and there is no significant difference between the mean errors in the topsoil and subsoil, which are very similar. The within-profile error variance is slightly larger than the corresponding variance component for the variable itself, but the other two variance components are notably smaller. The variance for the within-profile random effect on reparameterization is somewhat smaller than the corresponding error variance component (Fig. 3) suggesting that there was some scope for improving the parameters for application in the southern African domain. The scatter plot (Figure S3k) shows some dispersion around the bisector, but less of a shrinkage effect than for some of the other PTFs.

The PTF due to Oliveira et al. (2002), parameterized with data from the South American Tropics, has a significant positive bias, but error variance components all smaller than the corresponding ones for the variable itself. The scatterplot (Figure S3 g) shows some shrinkage in the predictions. The variance components for the random effects on reparameterization are not much different from the corresponding error variance components.

3.2. PTF predictions of θ_{-1500}

Fig. 6 shows the scatter plots of predicted VWC at −1500 kPa (θ_{-1500}) against the observed value for each of the PTFs considered. The PTFs of Botula et al. (2013), Dijkerman (1988), Lal (1978), Oliveira et al. (2002) and Van den Berg et al. (1997) for θ_{-1500} have the smallest sums of error variances (Fig. 7). For all the mean error falls in the

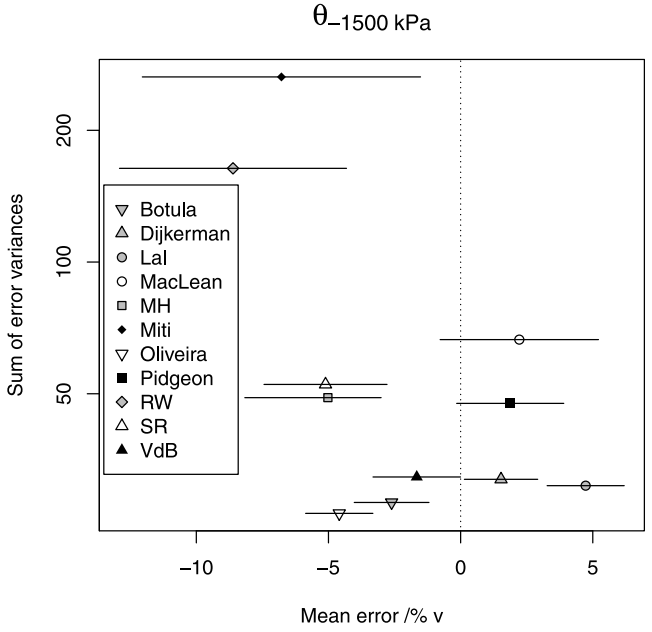


Fig. 7. For each PTF for θ_{-1500}, the estimate of mean error, with error bars (twice the standard error), plotted against the sum of the variance components for the error. The PTF authors in the key are as follows. Botula (Botula et al., 2012); Djikerman (Dijkerman, 1988); Lal (Lal, 1978); MacLean (MacLean and Yager, 1972); MH (Minasny and Hartemink, 2011); Miti (Miti, 2017); Oliviera (Oliveira et al., 2002); Pidgeon (Pidgeon, 1972); RW (Rawls and Brakensiek, 1982); SR (Saxton and Rawls, 2006); VdB (Van den Berg et al., 1997).

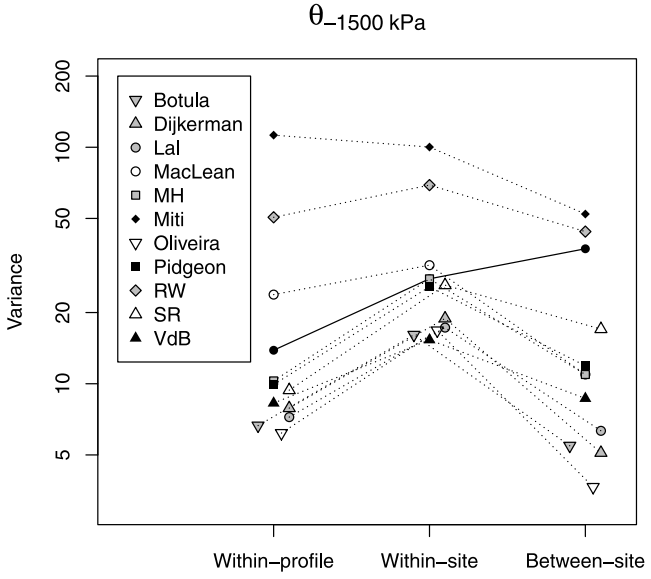


Fig. 8. Variances of the random effects for PTF error (θ_{-1500}), from the model with a constant mean as the fixed effects. The solid line with no symbols shows the corresponding variance components for θ_{-1500}. The PTF authors in the key are as follows. Botula (Botula et al., 2012); Djikerman (Dijkerman, 1988); Lal (Lal, 1978); MacLean (MacLean and Yager, 1972); MH (Minasny and Hartemink, 2011); Miti (Miti, 2017); Oliviera (Oliveira et al., 2002); Pidgeon (Pidgeon, 1972); RW (Rawls and Brakensiek, 1982); SR (Saxton and Rawls, 2006); VdB (Van den Berg et al., 1997).

interval [−5%,5%], and for none does the confidence interval include zero. The error variance components are all markedly smaller than the corresponding variance components of θ_{-1500} and are smallest at the between-site scale in all cases (Fig. 8). The variance components for the random effects on reparameterizing the PTFs of Botula et al. (2013), Dijkerman (1988) and Lal (1978) are all very similar to the corresponding error variance components (Fig. 9). For the PTF of Van den Berg et al. (1997) the within-profile variance for the reparameterized model is somewhat smaller than the corresponding error variance, for the PTF of Oliveira et al. (2002) the between-site variance component is somewhat larger than the corresponding error variance.

The predictors used in these five PTFs are not all the same. Dijkerman (1988) and Lal (1978) use clay content only, and these two PTFs show some degree of shrinkage, with under-prediction of larger values of θ_{-1500}. The PTFs of Van den Berg et al. (1997) and Botula et al. (2013) use equivalent predictor sets (two of sand, silt or clay, with bulk density). Both the scatter plots (Figures S7(k) and 6(a) respectively) show no systematic shrinkage effect, it is notable that these two PTFs are the ones with the smallest absolute overall mean error of prediction. The PTF of Oliveira et al. (2002) uses all three particle size fractions and bulk density. Given that the three fractions are a compositional variate, summing to 100%, this is not good practice. There is more shrinkage seen with this PTF (Figure S7(g)), with notable over-prediction of small values of θ_{-1500}.

For all this group of five PTFs, the difference in mean error between topsoil and subsoil is small, with the topsoil error slightly more positive (not a significant difference for the PTFs of Lal (1978) and Oliveira et al. (2002)). For the PTFs of Dijkerman (1988), Botula et al. (2013) and Van den Berg et al. (1997), it is notable from the exploratory plots (see the plot for Van den Berg et al., 1997 in Fig. 10), that there is some relation between the error and the clay or sand content, suggesting non-linear effects not fully captured in the PTF.

The PTFs of Saxton and Rawls (2006), Minasny and Hartemink (2011), and Pidgeon (1972) have somewhat larger sums of error variance components than do the five considered so far for this variable (Fig. 7). Their error variance components are all smaller than the corresponding components for θ_{-1500} (Fig. 8) and with the within-profile component the smallest and the within-site component the largest. Saxton and Rawls (2006) and Minasny and Hartemink (2011) have a negative mean error, significantly different from zero, and no significant difference in mean error between topsoil and subsoil (Figure S8). For all these PTFs the variance components for the reparameterized model are distinctly smaller than the corresponding error components (Fig. 9), most notably at the between-site level. It is notable that these PTFs all use soil organic matter or carbon as predictors, along with the proportion of one or more textural fractions. Exploratory plots for these PTFs are all suggestive of sources of error. For both Saxton and Rawls (2006) (Figure S9) and Minasny and Hartemink (2011) (Figure S10) there is a notable relation between PTF error and the clay or sand content (more negative errors for heavier soils) and bulk density (more negative errors for soils of smaller bulk density). For both Minasny and Hartemink (2011) and Pidgeon (1972) there is a clear negative bias introduced for soils with organic carbon content larger than 5% by mass, see the exploratory plot for Minasny and Hartemink (2011) in Figure S10. The scatterplots for these PTFs show no evidence of shrinkage effects. It is notable that there is a very strong cluster of points in the plot for Pidgeon (1972) (Figure S7(h)) parallel to the bisector, along with a "fringe" of over-predicted values.

MacLean and Yager's (1972) PTF for θ_{-1500} has a somewhat larger sum of error variances than the others considered so far (Fig. 7), and a positive mean error, although the confidence interval includes zero. The error variances at the within-profile and within-site levels are somewhat larger than the equivalent variances for the variable itself (Fig. 8). The variances for the random effects at these two levels are notably smaller for the reparameterization of the model than for the errors (Fig. 9). The exploratory plots for this PTF (Figure S11) show that soils with organic carbon content above 5% by mass tend to be markedly over-predicted for the variable, while many heavier-textured soils are under-predicted. It is notable that the range of concentrations of soil organic matter in the data set used by MacLean and Yager (1972) is much narrower than these data, with the maximum at less

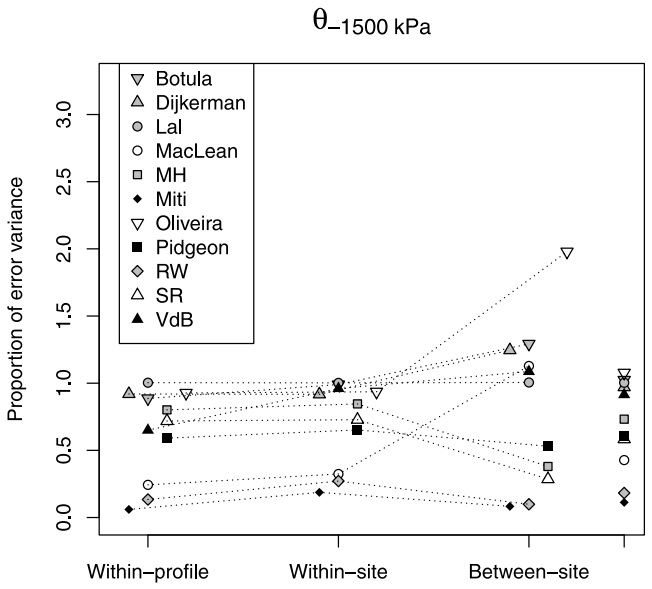


Fig. 9. Variances of the random effects from reparameterization of each θ_{-1500} PTF expressed as a proportion of the corresponding variance components for that PTF's error. The PTF authors in the key are as follows. Botula (Botula et al., 2012); Djikerman (Dijkerman, 1988); Lal (Lal, 1978); MacLean (MacLean and Yager, 1972); MH (Minasny and Hartemink, 2011); Miti (Miti, 2017); Oliviera (Oliveira et al., 2002); Pidgeon (Pidgeon, 1972); RW (Rawls and Brakensiek, 1982); SR (Saxton and Rawls, 2006); VdB (Van den Berg et al., 1997).

than 10%, close to equivalent to 5% SOC above which the bias is seen (Figure S2). Note that these effects are similar to those seen for Minasny and Hartemink (2011) and Pidgeon (1972), explaining the increased dispersion of the scatterplots for these PTFs with increasing θ_{-1500} (Figure S7d,e,h).

The PTFs of Miti (2017) and Rawls and Brakensiek (1982), have the largest sums of error variances (Fig. 7) and all three variance components for the error are larger than the corresponding components for the variable itself (Fig. 8). On reparameterization the variances of the random effects for these two PTFs are markedly smaller than the corresponding variances of the prediction error (Fig. 9). For both these PTFs there is a large negative mean error, with a larger absolute bias than for any other PTF for this property. In both cases the bias is more negative for the topsoil than the subsoil, and there is a significant depth effect (Figure S8). Both these PTFs generate unphysical values of θ_{-1500}, and inspection of the exploratory plots (Figure S12, for Rawls and Brakensiek, 1982), shows that in both there is a marked bias in predictions for soils with larger organic content, and in the case of Rawls and Brakensiek (1982), there is also a tendency to over-prediction of water content for heavier-textured soils. For both PTFs the organic matter content of the calibration data, and the clay contents, are rather narrower than for the validation data used here.

4. Discussion

4.1. PTF predictions of θ_{-33}

It is notable that the variance components for θ_{-33} at the within and between-site level are very similar, and larger than the within-profile variance (Fig. 3, solid discs joined by an unbroken line). For the PTFs with the four largest sum of error variance terms (as shown in Fig. 2), the error variances at the within-profile and within-site level are markedly larger than the between-site components. This may be because errors associated with the organic carbon or particle size effects have a big effect on predictions in contrasting horizons. For most of the PTFs in which the error variances are all smaller than (or close to) the variances of the original variable, the error variance at within-profile level is the smallest, and within-site the largest. These PTFs are particularly unsuitable for predicting variation of θ_{-33} within profiles. This is an important consideration for scientists who want to convert profile descriptions into information on vertical variation in soil water properties for modelling water and nutrient movement.

Many of the PTFs for θ_{-33} show a regression to the mean effect, with smaller values over-predicted and larger values under-predicted. This can be seen, for example, with the PTF of Lal (1978) in Fig. 1c. It is notable that this PTF uses only clay content as a predictor. This shrinkage effect should be of concern if PTFs are used to predict hydraulic properties of the soil for spatial models, because the tendency not to represent extremes may bias estimates of aggregate quantities made through non-linear process models, for example if one were to estimate total nutrient leaching for an aquifer or catchment.

Several of the PTFs produced unphysical predictions of θ_{-33}, which are associated with larger organic carbon concentrations. Although these errors cannot all be accounted for purely in terms of extrapolation outwith the range of the data used for calibration, as shown in Figure S2, this is likely to be a factor, e.g. for Miti (2017). This finding underlines the importance of careful and critical use of PTFs for modelling, particularly with large data sets. If PTFs are to be used routinely in workflows where their outputs (e.g. values of θ) are intermediate values only, and might not be directly inspected, then it is important to build in error-catching rules to identify PTF predictions which are implausible or unphysical.

The PTFs for θ_{-33} developed on temperate data from the US are not notably poorer than those developed in Africa or the Tropics more generally. Note, for example, the very similar scatterplots for the PTFs of Miti (2017), Figure S3(f), and Pidgeon (1972), Figure S3(h), PTFs from Zambia and Uganda respectively, and the second PTF for this tension from Rawls and Brakensiek (1982) in Figure S3(j). For the soils not obviously affected by the soil organic matter bias, these PTFs all show marked shrinkage in the predictions (i.e. less variation of the predicted values than the corresponding observations). The PTF due to Saxton and Rawls (2006) shows little shrinkage effect, Figure S3(k). This shows that a PTF selected from the literature should not necessarily be accepted or rejected for use in a particular study simply on the basis of the geographical domain in which it was developed and in which it is to be applied. However, a comparison of the range of basic soil properties in the two domains may well be useful for decisions on suitability.

Our evaluation of a set of PTFs from the literature allows us to draw some general conclusions about their possible limitations. They also allow us to make some recommendations for the use of PTFs to predict θ_{-33} for soils in Zimbabwe, Zambia and Malawi with the range of values of basic properties in our legacy data set.

1. In terms of bias and prediction error variances, the PTF of Botula et al. (2013) for θ_{-33} appears to perform well, with only small changes on reparameterization. None the less, it does show some shrinkage, which could reduce its usefulness in some contexts
2. The PTF of MacLean and Yager (1972) for θ_{-33} has marked bias, but the error variance components are small. There are marked changes on reparameterization, but this suggests that the assemblage of predictors could be useful in the region. Note this is the only PTF to use coarse sand as a predictor.
3. The first PTF of Rawls and Brakensiek (1982) for θ_{-33} shows very little shrinkage, and may well be useful for soils with organic carbon less than 5% by mass.
4. The PTF of Saxton and Rawls (2006) for θ_{-33}, which combines polynomial terms in the predictors, has a small bias, but shows little shrinkage, has a relatively small sum of error variance components and might also be useful in this setting.

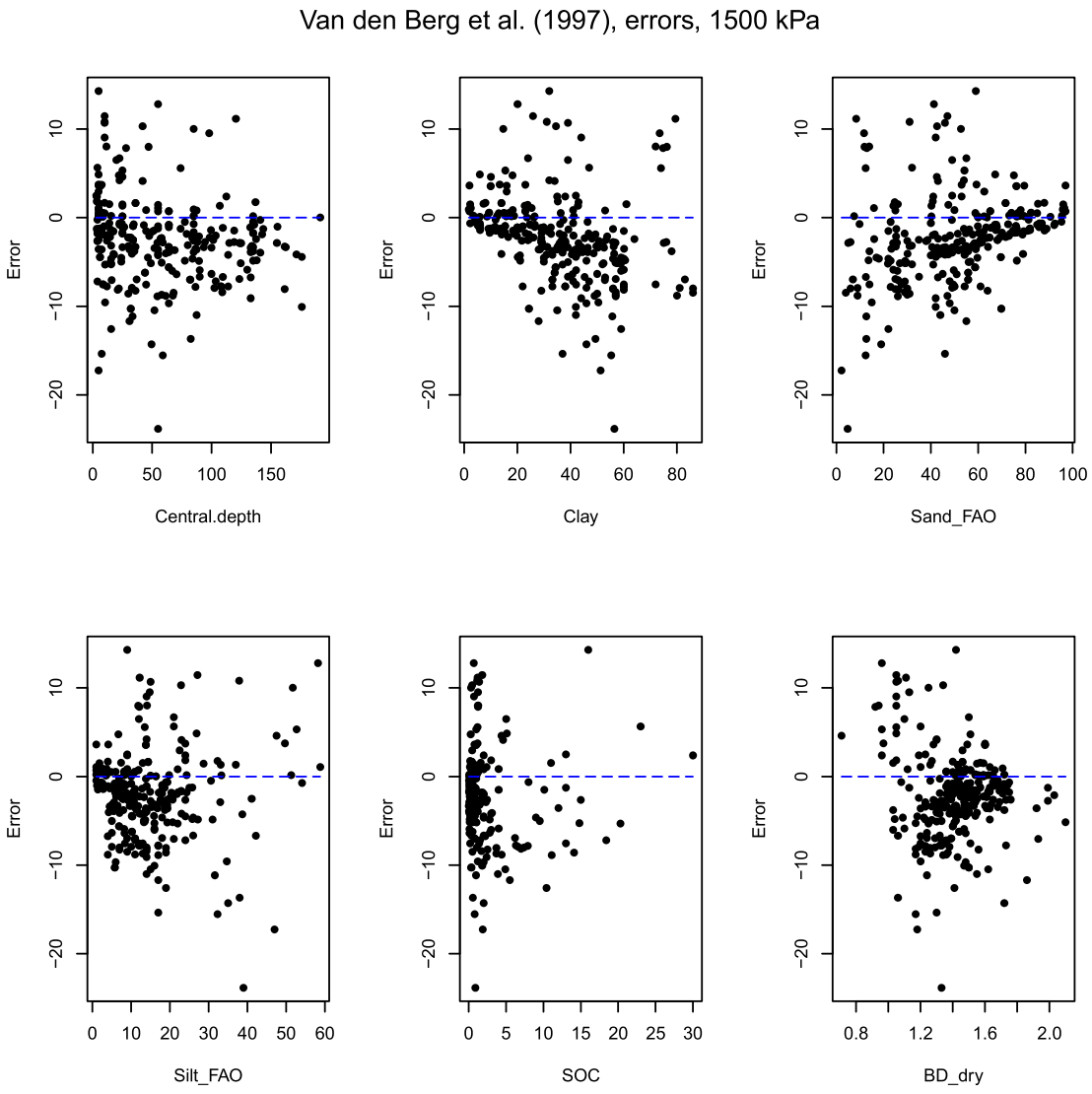


Fig. 10. Scatter plots of PTF error against the central depth of the observation, clay content, sand content, silt content, SOC content and dry bulk density. This is for the PTF for θ_{-1500} of Van den Berg et al. (1997).

4.2. PTF predictions of θ_{-1500}

There are two distinct groups of PTFs for θ_{-1500} with respect to the performance of their predictions. PTFs in the first group (Botula et al., 2013; Dijkerman, 1988; Lal, 1978; Oliveira et al., 2002; Van den Berg et al., 1997), have the smallest error variances and, on reparameterization, the variance components for the random effects are close to the corresponding variances in the model for PTF errors. The mean bias for these PTFs are in the range $[-5\%, 5\%]$. Of these five PTFs, three have a calibration domain in tropical Africa and two are in the wider Tropics. The PTFs due to Saxton and Rawls (2006), Minasny and Hartemink (2011) and Pidgeon (1972) (one from a temperate domain, on generic Tropical PTF, and one from an African domain respectively) show somewhat larger error variances and reparameterization effects, and have clear errors due to particle size or SOC effects. Similar, but more pronounced effects are seen for the PTFs of Miti (2017) and Rawls and Brakensiek (1982). It is notable that PTFs show less shrinkage in the predicted values for θ_{-1500} than we observed for θ_{-33}.

For θ_{-1500} the best-performing PTFs are from tropical settings, but again it is clear that this does not guarantee good performance. The range of values for the predictor variables must be checked, and, as for θ_{-33}, it is necessary to check for implausible or unphysical results. In the absence of other information one might therefore select one of the first set of PTFs to predict θ_{-1500} in the southern African setting.

5. Conclusions

Differences were found in the performance of published PTFs when they were tested on a set of legacy data from Zambia, Zimbabwe and Malawi. The prediction of θ_{-33} and θ_{-1500} from soil properties measured in soil surveys could be undertaken with selected PTFs from the examined set. It was notable, in the case of θ_{-33}, that PTFs calibrated at other African or tropical sites were not necessarily the best, but it was clearly important to ensure that the range of predictor values in the calibration data set was comparable with the range of values over which the PTF was to be applied. However, PTFs originally calibrated with data from an African or other tropical calibration domain were notably the best for prediction of θ_{-1500}.

The legacy data used in this study were strongly clustered in space with multiple observations in different horizons of common soil profiles, and multiple profiles observed at the same site. We contend that this state of affairs is likely to be typical, as many legacy data sources are soil survey reports which include profile descriptions, and were several profiles may be examined over a surveyed farm or estate. Our evaluation of the PTFs was therefore based on linear mixed models

to account for correlations within-profile, within-site (between profile) and spatially dependent between-site correlation. The estimated random effects were very informative about the scales at which published PTFs are most reliable, with error variance typically small relative to the native variance of the variables at the between-site scale. Most of the PTFs observed here show the largest reduction in unexplained variation at the between-cluster scale. This means that they will be particularly useful for modelling large-scale variations in soil and crop behaviour. However, they may be less reliable for predicting variation at within-farm or within-field scale (as might be required for precision management of inputs) or at within-profile scale (as might be required for modelling water and nutrient movement in the rooting zone). Not all studies which use legacy data on soil pay attention to the structure of the data and its likely implications, e.g. Hengl et al. (2021)

We propose that such studies, based on legacy data, and with a suitable LMM, should be used to screen PTFs before their wider application, and, where possible, to reestimate them for local use.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgements

This work was supported by UK Research and Innovation (UKRI) [grant number NE/P02095X/1] through the GROW programme of the Global Challenges Research Fund. The project is entitled *CEPHaS - Strengthening Capacity in Environmental Physics, Hydrology and Statistics for Conservation Agriculture Research*.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geoderma.2023.116661.

References

Aina, P., Periaswamy, S., 1985. Estimating available water-holding capacity of western Nigerian soils from soil texture and bulk density, using core and sieved samples. Soil Sci. 140 (1), 55–58.
Batjes, N., 1996. Development of a world data set of soil water retention properties using pedotransfer rules. Geoderma 71 (1–2), 31–52.
Batjes, N., Ribeiro, E., Van Oostrum, A., Leenaars, J., Hengl, T., Mendes de Jesus, J., 2017. WoSIS: Providing standardised soil profile data for the world. Earth Syst. Sci. Data 9 (1), 1–14.
Botula, Y., Cornelis, W., Baert, G., Van Ranst, E., 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (DR Congo). Agricult. Water Manag. 111, 1–10.
Botula, Y., Nemes, A., Mafuka, P., Van Ranst, E., Cornelis, W., 2013. Prediction of water retention of soils from the humid tropics by the nonparametric-nearest neighbor approach, vadose zo. j., 12, 1–17.
Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16 (5), 1190–1208.
Diggle, P.J., Ribeiro, P.J., 2007. An overview of model-based geostatistics. Model-Based Geostat. 27–45.
Dijkerman, J., 1988. An ustult-aquult-tropept catena in Sierra Leone, West Africa, II. Land qualities and land evaluation. Geoderma 42 (1), 29–49.
Givi, J., Prasher, S., Patel, R., 2004. Evaluation of pedotransfer functions in predicting the soil water contents at field capacity and wilting point. Agricult. Water Manag. 70 (2), 83–96.
Gneiting, T., 2013. Strictly and non-strictly positive definite functions on spheres. Bernoulli 19 (4), 1327–1349.
Gupta, S., Larson, W., 1979. Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. Water Resour. Res. 15 (6), 1633–1635.
Hall, D., Reeve, M., Thomasson, A., Wright, V., 1977. Water Retention, Porosity and Density of Field Soils. Tech. Rep..
Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., et al., 2021. Introducing AfroGrid, a unified framework for environmental conflict research in Africa. Sci. Rep. 11, 6130.
Hijmans, R.J., Williams, E., Vennes, C., 2017. Geosphere: Spherical trigonometry. R package version 1.5-7.
IUSS Working Group WRB, 2006. World Reference Base for Soil Resources 2006, World Soil Resources Report No. 103. Food and Agriculture Organization of the United Nations, Rome.
Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Micheli, E., Montanarella, L., Spaargaren, O., Tahar, G., Thiombiano, L., Van Ranst, E., Yemefack, M., Zougmore, R.E., 2013. Soil Atlas of Africa. European Commission. Publication Office of the European Union, Luxembourg.
Lal, R., 1978. Physical properties and moisture retention characteristics of some Nigerian soils. Geoderma 21 (3), 209–223.
Landon, J.R., 2014. Booker Tropical Soil Manual: A Handbook for Soil Survey and Agricultural Land Evaluation in the Tropics and Subtropics. Routledge.
Lark, R., Cullis, B., Welham, S., 2006. On spatial prediction of soil properties in the presence of a spatial trend: The empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 57 (6), 787–799.
MacLean, A.H., Yager, T.U., 1972. Available water capacities of Zambian soils in relation to pressure plate measurements and particle size analysis. Soil Sci. 113 (1), 23–29.
Minasny, B., Hartemink, A.E., 2011. Predicting soil properties in the tropics. Earth-Sci. Rev. 106 (1–2), 52–62.
Miti, C., 2017. Prediction of Soil Physical and Hydraulic Properties of Zambian Soil (Master dissertation thesis). Ghent University, Gent, Belgium,Gent University.
Mugabe, F.T., 2004. Pedotransfer functions for predicting three points on the moisture characteristic curve of a Zimbabwean soil. Asian Journal of Plant Science 3 (6), 679–682.
Nemes, A., Roberts, R., Rawls, W.J., Pachepsky, Y.A., Van Genuchten, M.T., 2008. Software to estimate- 33 and- 1500 kPa soil water retention using the non-parametric k-nearest neighbor technique. Environ. Model. Softw. 23 (2), 254–255.
Oliveira, L., Ribeiro, M., Jacomine, P., Rodrigues, J., Marques, F., 2002. Pedotransfer functions for the prediction of moisture retention and specific potentials in soils of Pernambuco State (Brazil). Rev. Brasileira de Ciência do Solo 26, 315–323.
Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58 (3), 545–554.
Pidgeon, J., 1972. The measurement and prediction of available water capacity of ferrallitic soils in Uganda. J. Soil Sci. 23 (4), 431–441.
R Core Team, 2020. R: A language and environment for statistical computing. URL https://www.R-project.org/.
Rawls, W.J., Brakensiek, D., 1982. Estimating soil water retention from soil properties. J. Irrigation Drain. Div. 108 (2), 166–171.
Rawls, W.J., Brakensiek, D.L., 1985. Prediction of soil water properties for hydrologic modeling. In: Watershed Management in the Eighties. ASCE, pp. 293–299.
Salter, P., Haworth, F., 1961. The available-water capacity of a sandy loam soil: I. a critical comparison of methods of determining the moisture content of soil at field capacity and at the permanent wilting percentage. J. Soil Sci. 12 (2), 326–334.
Saxton, K.E., Rawls, W.J., 2006. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. Soil Sci. Soc. America J. 70 (5), 1569–1578.
Saxton, K.E., Rawls, W.J., Romberger, J.S., Papendick, R.I., 1986. Estimating generalized soil-water characteristics from texture. Soil Sci. Soc. America J. 50 (4), 1031–1036.
Schaap, M.G., Leij, F.J., Van Genuchten, M.T., 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251 (3–4), 163–176.
Schon, J., Koren, I., 2022. Introducing AfroGrid, a unified framework for environmental conflict research in Africa. Sci. Data 9, 116.
Shein, E., Arkhangel'Skaya, T., 2006. Pedotransfer functions: State of the art, problems, and outlooks. Eurasian Soil Sci. 39 (10), 1089–1099.
Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media.
Tomasella, J., Hodnett, M., 2004. Pedotransfer functions for tropical soils. Dev. Soil Sci. 30, 415–429.
Van den Berg, M., Klamt, E., Van Reeuwijk, L., Sombroek, W., 1997. Pedotransfer functions for the estimation of moisture retention characteristics of Ferralsols and related soils. Geoderma 78 (3–4), 161–180.
Verbeke, G., 1997. Linear mixed models for longitudinal data. In: Linear Mixed Models in Practice. Springer, pp. 63–153.
Vereecken, H., Maes, J., Feyen, J., Darius, P., 1989. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. Soil Sci. 148 (6), 389–403.
Wösten, J., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database of hydraulic properties of European soils. Geoderma 90 (3–4), 169–185.
Wösten, J., Verzandvoort, S., Leenaars, J., Hoogland, T., Wesseling, J., 2013. Soil hydraulic information for river basin studies in semi-arid regions. Geoderma 195, 79–86.