Received: 29 March 2021 | Revised: 22 December 2021 | Accepted: 25 January 2022

DOI: 10.1111/ejss.13220

ORIGINAL ARTICLE

European Journal of Soil Science WILEY

Modelling soil water retention and water-holding capacity with visible–near-infrared spectra and machine learning

Philipp Baumann[1,2] | **Juhwan Lee**[3,4] | **Thorsten Behrens**[5] | **Asim Biswas**[6] | **Johan Six**[2] | **Gordon McLachlan**[7] | **Raphael A. Viscarra Rossel**[1,3]

[1]CSIRO Land & Water, Bruce E. Butler Laboratory, Canberra, Australia

[2]Institute of Agricultural Sciences, Department of Environmental Systems Science, Swiss Federal Institute of Technology, ETH Zurich, Zurich, Switzerland

[3]Soil & Landscape Science, School of Molecular and Life Sciences, Curtin University, Perth, Australia

[4]Department of Smart Agro-industry, Gyeongsang National University, Jinju, Republic of Korea

[5]Soil and Spatial Data Science, Solution GbR, Quedlinburg, Germany

[6]School of Environmental Sciences, University of Guelph, School of Environmental Sciences, Guelph, Canada

[7]CSIRO Agriculture, Canberra, Australia

Correspondence
Raphael A. Viscarra Rossel, Soil & Landscape Science, School of Molecular and Life Sciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia.
Email: r.viscarra-rossel@curtin.edu.au

Funding information
Grains Research and Development Corporation

Abstract

We need measurements of soil water retention (SWR) and available water capacity (AWC) to assess and model soil functions, but methods are time-consuming and expensive. Our aim here was to investigate the modelling of AWC and SWR with visible–near-infrared spectra (vis–NIR) and the machine-learning method CUBIST. We used soils from 54 locations across Australian agricultural regions, from three depths: 0–15 cm, 15–30 cm and 30–60 cm. The volumetric water content of the samples and their vis–NIR spectra were measured at seven matric potentials from −1 kPa to −1500 kPa. We modelled the following: (i) AWC directly with the average spectra of the samples measured at different water contents, (ii) water contents at field capacity and permanent wilting point and calculated AWC from those estimates, (iii) AWC with spectra of air-dried soils, and (iv) parameters of the Kosugi and van Genuchten SWR models, then reconstructed the SWR curves to calculate AWC. We compared the estimates with those from a local pedotransfer function (PTF) and an established Australian PTF. The accuracy of the spectroscopic approaches varied but was generally better than the PTFs. The spectroscopic methods are also more practical because they do not require additional soil properties for the modelling. The root-mean squared-error (RMSE) of the spectroscopic methods ranged from 0.033 $cm^3\ cm^{-3}$ to 0.059 $cm^3\ cm^{-3}$. The RMSEs of the PTFs were 0.050 $cm^3\ cm^{-3}$ for the local and 0.077 $cm^3\ cm^{-3}$ for the general PTF. Spectroscopy with machine learning provides a rapid and versatile method for estimating the AWC and SWR characteristics of diverse agricultural soils.

Highlights

- Soil available water capacity can be estimated with vis-NIR specta.
- Parameters of water retention models can be estimated with vis-NIR spectra.
- vis-NIR spectroscopy performed better than pedotransfer functions.
- The results apply to a diverse range of soils.

KEYWORDS
available soil water, machine learning, soil water retention, visible–near-infrared spectroscopy, water retention models

1 | INTRODUCTION

Water sustains life in the soil and supports the functioning of ecosystems. The estimation of water contents at different matric potentials and the available water holding capacity (AWC) of soils are needed to describe the water status in a soil–plant system and are important model inputs for simulating soil–plant processes. Information on the amount of water available to plants is essential for understanding plant growth, development, and physiology (Passioura, 2002). Soil water also affects organic matter decomposition, soil microbial activity and the physical and biological conditions of the soil (Manzoni et al., 2012; Skopp et al., 1990). Soil water is vital for agricultural productivity and management, for example, to improve irrigation and fertiliser-use efficiencies.

Soil water retention (SWR) characteristics, which can be expressed with parametric models (e.g., Kosugi (1994) and van Genuchten (1980)), are used to understand and predict water and solute transport in variably saturated soils and the exchange of gases and vapour between the soil and the atmosphere (Vereecken et al., 2016). Despite the importance of SWR and the AWC, however, there are few new practical and cost-efficient methods for measuring them. Vereecken et al. (2008) reported on a range of methods for measuring soil water at different scales, which include ground-based (proximal), wireless sensor networks, and airborne (remote) sensors.

Pedotransfer functions (PTF) (Bouma, 1989) are often used to estimate AWC and SWR using more readily available soil physical and chemical properties (e.g., clay content, bulk density and organic matter). Over the last three decades, there has been much research to develop and improve PTFs using different statistical methods, including artificial neural network, regression trees, k-nearest neighbours, support vector machine and genetic programming (Minasny et al., 1999; Pachepsky & Rawls, 2004; Shein & Arkhangel'skaya, 2006; Van Looy et al., 2017; Wösten et al., 2001). One advantage of PTFs is that they are easy to develop and implement if the soil properties needed to parametrize them are already available. However, to derive PTFs for site-specific, local application, measurements of other soil properties are required, increasing the cost of the survey, or one must rely on general, historical data to develop them. Estimates made with general PTFs can be very inaccurate because of the differences in soil types, the inevitable extrapolations due to inadequate feature space coverage and the possible mismatch of scale (Pachepsky & Hill, 2017; Van Looy et al., 2017).

Soil spectroscopy in the visible–near-infrared (vis–NIR; 400–2500 nm) can quantify absorptions that result from the interaction between chemical bonds and photons, causing energy-specific vibrations of chemical bonds present in water, organic matter and mineral constituents in soils. A particular chemical bond typically has several fundamental vibrations in the mid-infrared, which also occur in the NIR, but as weaker and broader overtones and combination bands. Most soil properties result from the inherent composition of the soil: namely, its minerals (such as clay minerals, iron oxides, quartz), its organic matter, water, and air. Thus, soil spectra can provide an integrative, multivariate measure of soil composition and can be used to represent the soil's physicochemical and biological characteristics. Over the last two decades, soil spectroscopy, combined with multivariate statistics and machine learning, has provided accurate estimates of various soil properties, including soil water content (Soriano-Disla et al., 2014; Stenberg et al., 2010; Viscarra Rossel et al., 2016).

Research on the relationship between water content and vis–NIR spectra is abundant (Bowers & Hanks, 1965; Curcio & Petty, 1951; Lobell & Asner, 2002; Stoner et al., 1980; Viscarra Rossel & McBratney, 1998). Soil vis–NIR spectra show absorptions of hydroxyl bonds in water vibrating at specific wavelengths. Absorptions that have been used to estimate soil water in the vis–NIR are those near 450, 600, 1200, 1400, 1900, 2100, and 2400 nm (Ben-Dor et al., 1999; Bishop, 1994; Knadel et al., 2014; Soltani et al., 2018; Weidong et al., 2002; Whiting et al., 2004). However, there are fewer studies on the modelling of AWC and SWR with vis–NIR spectra, and some perform the modelling using vis–NIR spectra with only air-dry soils (Babaeian et al., 2015,b; Blaschek et al., 2019; Knadel et al., 2014; Pittaki-Chrysodonta et al., 2018; Santra et al., 2009). There are no studies that provide a comprehensive assessment of the potential for soil vis–NIR spectroscopy to estimate AWC and SWR over a large extent or with a diverse set of soils. Thus, our aim here is to (i) explore the relationship between soil water retention and vis–NIR spectra and (ii) investigate different approaches for spectroscopic modelling the AWC and SWR characteristics of a diverse set of soils from across the main agricultural regions in Australia using the machine learning algorithm CUBIST.

2 | MATERIALS AND METHODS

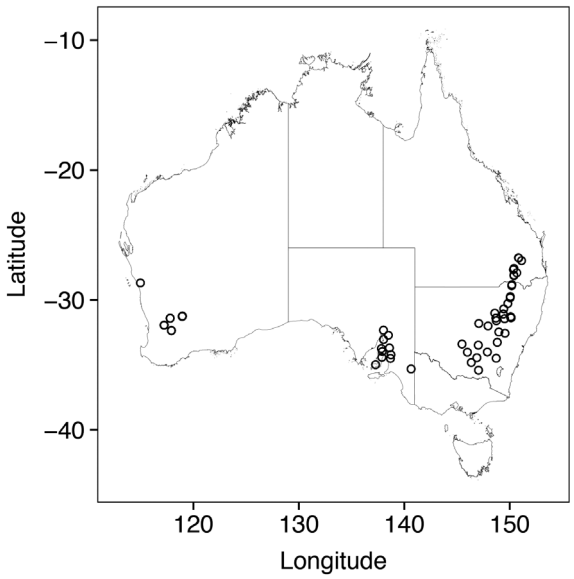
2.1 | Soil sampling, laboratory analyses and spectroscopy

We performed experiments to derive SWR curves for a range of Australian agricultural soils sampled from three depths: 0–15 cm, 15–30 cm and 30–60 cm. We acquired the soil samples from the Agricultural Production Systems Research Unit (APSRU) and the Commonwealth Scientific and Industrial Research Organisation's (CSIRO) National Soil Archive. They originate from 54 locations, which cover a vast geographic extent across the Australian wheat-sheep belt and represent a diverse set of soil types (Figure 1).

The archived soil samples were crushed and sieved to a particle size of ≤2 mm. For our experiments, we packed the samples into polyvinyl chloride cylinders (40 mm in diameter and 30 mm height). We measured the bulk density of the soil samples, and their volumetric water contents at seven matric potentials: −1 kPa, −5 kPa, −10 kPa, −29 kPa (field capacity; FC) and − 60 kPa using suction plates, − 500 kPa and −1500 kPa (permanent wilting point; PWP) in a pressure chamber, and at air-dry conditions. We used the common definition of FC at $\psi = -29$ kPa, although in Australia the common definition is $\psi = -10$ kPa. This was done because the Australian convention is guided toward sandy soils, but we had a significant range in textures. After the measurements at each matric potential, we recorded the vis–NIR spectra of the soils from both the top and bottom of the cylinders to obtain a representative measure (see below). The packed soils were subsampled and then oven-dried over-night at 60°C, weighed, ground, and stored at room temperature until soil analyses. We measured the organic carbon content of the soil samples using the dry combustion method (Rayment & Higginson, 1992) and their clay, sand and silt fractions with the hydrometer method (Gee & Bauder, 1986). We used texture classes commonly used in Australia, which are clay <2 μm; silt 2–20 μm and sand 20–2000 μm (Bowman & Hutka, 2002).

We recorded the vis–NIR spectra of the samples with a Labspec® vis–NIR spectrometer (PANalytical Inc., Boulder, CO, USA, formerly Analytical Spectral Devices) with a spectral range of 350–2500 nm and spectral resolution of 3 nm at 700 nm and 10 nm at 1400 and 2100 nm. Measurements were made with a high-intensity contact probe illuminated by a halogen bulb (2901 ± 10 K). The contact probe measures a spot of roughly 10 mm diameter and is designed to minimise errors associated with stray light. We calibrated the sensor with a Spectralon® (Labsphere, North Sutton, New Hampshire, USA) white reference panel once every ten measurements to account for changing laboratory conditions (e.g., humidity, temperature). The sampling spectral resolution of the spectrometer was 1 nm so each spectrum was comprised of 2151 wavelengths. Measurements were made following protocols described in Viscarra Rossel et al. (2016). The soils were measured in quadruplicate, with two spectra recorded on the upper surface and two on the lower surface of each core. We averaged these replicates to produce a single measurement per sample, thus, there were a total of 1296 spectra (54 locations × 3 depths × 8 water contents–7 matric potentials and air-dry).

To standardise the spectra for further analyses, we subtracted the reflectance of the first wavelength (with the minimum reflectance value) to correct for potential baseline shifts between the measurements. Because the



Soil order (ASC)	Locations
Dermosols (DE)	13
Chromosols (CH)	12
Sodosols (SO)	8
Kandosols (KA)	6
Calcarosols (CA)	5
Vertosols (VE)	4
Kurosols (KU)	3
Ferrosols (FE)	1
Podosols (PO)	1
Tenosols (TE)	1

FIGURE 1 Locations of the 54 soil sampling sites in Australian agricultural regions and the different Australian soil classification orders (Isbell, 2002) that they represent

spectra are highly collinear, we retained only every 10th wavelength from 350 to 2500 nm, resulting in 216 wavelengths for each sample. We used the continuum removed spectra (Clark & Roush, 1984) to enhance and compare characteristic absorptions at different matric potentials. For the spectroscopic modelling (see below), we used the Savitzky–Golay first derivative spectra with a cubic polynomial and window size of 13 points (Savitzky & Golay, 1964).

2.2 | Estimation of available water capacity

We calculated AWC of the soil samples by $AWC = \theta_{FC} - \theta_{PWP}$, where θ_{FC} and θ_{PWP} are the volumetric water contents, θ, measured at field capacity (FC; $\psi_{FC} = -29$ kPa) and permanent wilting point (PWP; $\psi_{PWP} = -1500$ kPa), respectively. We then modelled the AWC of the soils with the spectra and the machine learning method CUBIST (see Section 2.4). We modelled with four spectroscopic approaches and two PTFs, one derived locally and a general PTF. These experiments are described below and summarised in Figure 2.

In the first approach (A: SPC-AVG), we modelled AWC directly with the average spectra of the measurements between matric potentials $\psi = -1$ kPa and -1500 kPa. In the second approach (B: SPC-FC-PWP), we first estimated water contents at field capacity ($\psi = -29$ kPa [$\widehat{\theta_{FC}}$]) and at permanent wilting point (-1500 kPa [$\widehat{\theta_{PWP}}$]), by modelling their corresponding measured θ with the spectra and then using the estimates we calculated AWC. In the third approach (C: SPC-DRY), we modelled AWC using the air-dry spectra. In approach (D: SPC-SWRC), we fitted the measured data with the Kosugi (Kosugi, 1994) and van Genuchten SWR models (van Genuchten, 1980) using nonlinear least-squares optimization (see Section 2.3). We then modelled the fitted parameters of each model with the average spectra over all ψ. The estimated parameters from each SWR model were used to reconstruct the SWR curves and to estimate $\widehat{\theta_{FC}}$ and $\widehat{\theta_{PWP}}$ to then calculate AWC. We developed the local PTF, (E:

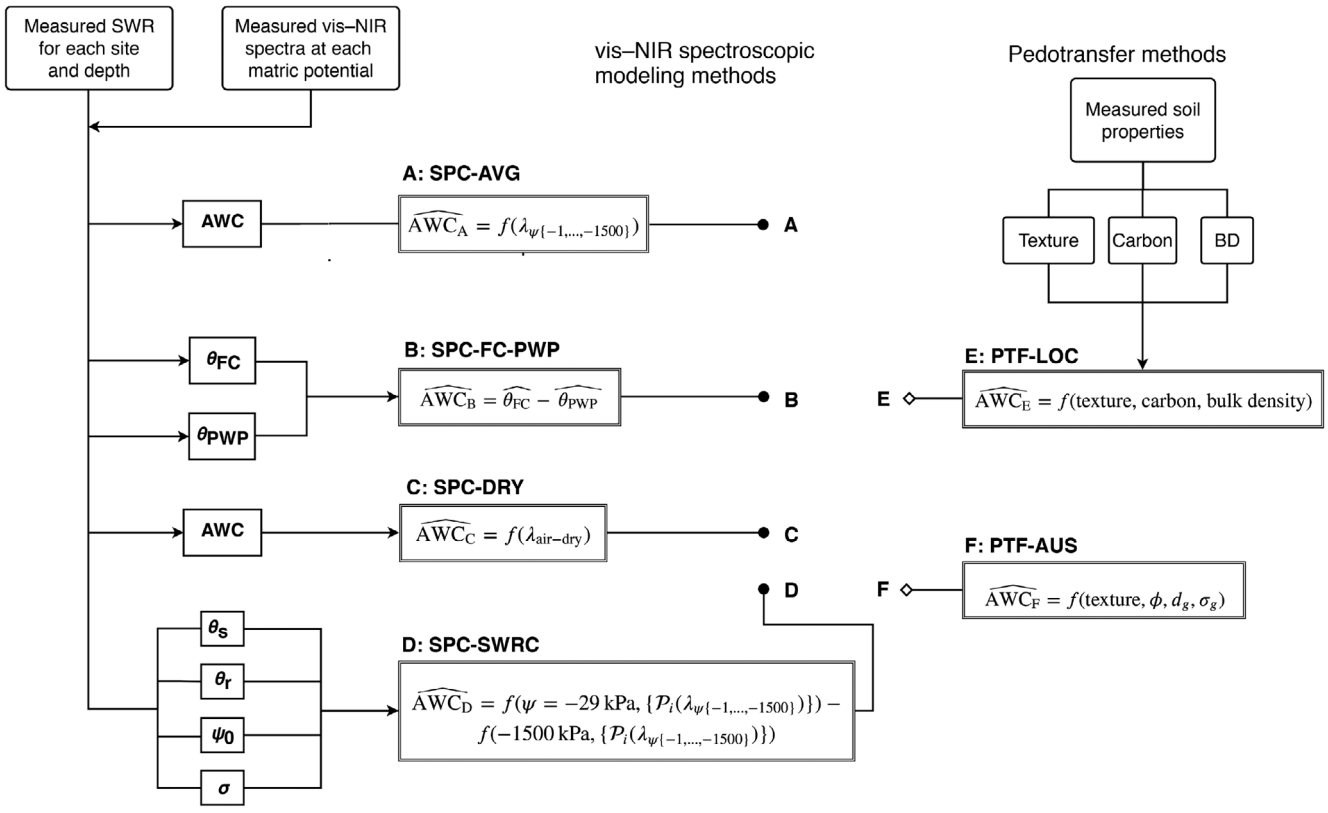


FIGURE 2 Different approaches to estimate the available water capacity (AWC) of soils: (A: SPC-AVG) direct spectroscopic modelling of AWC, (B: SPC-FC-PWP) modelling water contents at permanent wilting point (θ_{PWP}; $\psi = -1500$ *kPa*) and at field capacity (θ_{FC}; $\psi = -29$ *kPa*), then using those estimates to calculate AWC, (C: SPC-DRY) spectroscopic modelling using spectra of air-dry soils, (D: SPC-SWRC) spectroscopic modelling of the parameters of a soil water retention (SWR) model (here, the Kosugi and van Genuchten models) and estimating AWC from the reconstructed curve, (E: PTF-LOC) estimating AWC with a local pedotransfer (PTF) function developed with this study's data, and (F: PTF-AUS) prediction of AWC using the PTF developed by Minasny et al. (1999). λ represents the wavelengths in the spectra, BD is bulk density, ϕ is porosity, d_g is the geometric mean of the particle size diameters, σ_g is the geometric standard deviation of the particle size diameters

PTF-LOC), with measurements of soil texture, total organic carbon and bulk density and using CUBIST (see Section 2.4). The general one, (F: PTF-AUS), was derived using the function in Minasny et al. (1999) (see Data S1).

2.3 | Soil water retention models, fitting and optimisation

We characterised the SWR of the samples using the Kosugi (1994) and van Genuchten (1980) models. We used the two-parameter variant of the Kosugi model (not the full three-parameter model, which assumes no air-entry value).

2.3.1 | The Kosugi model

Kosugi (1994) proposed a soil water retention (SWR) model, assuming the pore radii to be log-normally distributed. The model describes the effective saturation S_e:

$$S_e = 0.5\ \mathrm{erfc}\left[\frac{\left(\ln\left\{\frac{\psi_c\text{-}\psi}{\psi_c\text{-}\psi_0}\right\}\text{-}\sigma^2\right)}{\sqrt{2}\sigma}\right] \quad \psi < \psi_c \qquad (1)$$

$$S_e = 1 \quad \psi \geq \psi_c$$

where erfc is the complementary error function, ψ_c is the air-entry value or the bubbling pressure, ψ_0 the inflection point, and $\sigma(\sigma > 0)$ is a dimensionless parameter related to the width of the pore radius distribution function $g(r)$ or the standard deviation of the log-transformed soil pore radius (r), given by:

$$g(r) = \frac{\theta_s - \theta_r}{(2\pi)^{1/2}\sigma r}\exp\left\{-\frac{[\ln r/r_m]^2}{2\sigma^2}\right\} \qquad (2)$$

where r_m is the median and geometric mean of $g(r)$, and θ_s and θ_r are saturated and residual water contents, respectively. Generally, the smaller the σ value is, the steeper becomes the retention curve at the inflection point (ψ_0).

When assuming no air-entry value, given by $\psi_c = 0$, the full three-parameter Kosugi model in Equation (1) simplifies to its two-parameter form.

2.3.2 | The van Genuchten model

We also used the van Genuchten (1980) model in our experiments because it is one of the most versatile and commonly used SWR models. The van Genuchten model was fitted to the volumetric water content measurements using:

$$S_e = [1 + (\alpha\psi)^n]^{-m} \qquad (3)$$

where S_e is the effective saturation or normalised water content, ψ the matric potential, α is a parameter inversely related to the air entry value, which is the matric potential where the largest connected soil pores are filled with air, n is a dimensionless measure of pore size distribution, and m is a dimensionless parameter related to n:

$$m = 1 - \frac{1}{n}\ (n > 1, 0 < m < 1). \qquad (4)$$

2.3.3 | Fitting and optimisation

We used the Levenberg–Marquardt non-linear least-squares (NLS) method (Marquardt, 1963) to fit the Kosugi and Van Genuchten models to the seven matric potentials measured at each site and depth (see above). Saturated and residual water contents θ_s and θ_r were used as two of the four fitting parameters of each model since most SWR models describe water retention in the range $\theta_r \leq \theta \leq \theta_s$. Parameter estimates are often sensitive to the chosen starting values because single parameter starting values can find local minima instead of global optima. Therefore, to prevent unrealistic parameter optimization, grids of 1000 candidate starting value combinations were randomly drawn from a uniform distribution from the upper and lower bounds of each parameter and site-depth combination. For both Kosugi and van Genuchten models, the upper and lower bounds for the starting parameters of θ_s were set to be 0.90 and 0.35 $cm^3\ cm^{-3}$, and the bounds for θ_r were 0.25 and 0.02 $cm^3\ cm^{-3}$. The selected θ_s and θ_r parameters were allowed to be within ±10% of the range of the measured water contents. For the Kosugi model, the upper and lower bounds of σ were 0.01 and 10 and for ψ_0 they were −98 kPa and −0.098 kPa. For the van Genuchten model, the bounds for α were 0.005 and 3 cm^{-1} and for n 1 and 15. The final, best-fit parameters for each SWR curve were selected using the Akaike information criterion (AIC) (Akaike, 1973), implemented in the nls.multstart R library (Padfield & Matheson, 2020), which we used. The AIC is defined by: AIC = $2k + n\ \ln(\mathrm{RSS}) - (n\ \ln[n] + 2C)$ where n is the number of data, C is a constant, k is the number of parameters for each SWR curve, and RSS is the residual sums of squares. We note that the AIC is generally used to compare different models with different numbers of parameter. In this case, however, the number of parameters is constant so that the AIC and the

Soil property	n	Mean	SD	Minimum	Median	Maximum
Bulk density [g cm^{-3}]	148	1.21	0.15	0.92	1.18	1.67
Total C [g kg^{-1}]	148	0.73	0.47	0.05	0.65	3.33
Clay [%]	148	37.0	16.0	3.0	37.5	82.0
Silt [%]	148	12.4	6.0	0.4	12.0	33.0
Sand [%]	148	50.4	17.9	9.0	48.5	96.0
θ_{FC}[cm^3 cm^{-3}]	162	0.291	0.135	0.062	0.279	0.665
θ_{PWP}[cm^3 cm^{-3}]	162	0.147	0.083	0.005	0.151	0.350
AWC [cm^3 cm^{-3}]	162	0.145	0.060	0.034	0.136	0.330

Note: θ_{FC} at $\psi = -29$ kPa and θ_{PWP} at $\psi = -1500$ kPa denote volumetric water contents at field capacity (FC) and permanent wilting point (PWP). The available water capacity (AWC) was calculated from θ_{FC} and θ_{PWP}. For some of the soil properties, $n = 14$ data were missing.

TABLE 1 Summary statistics of soil properties measured at 54 locations and 3 depths (0–15 cm, 15–30 cm and 30–60 cm) across Australia

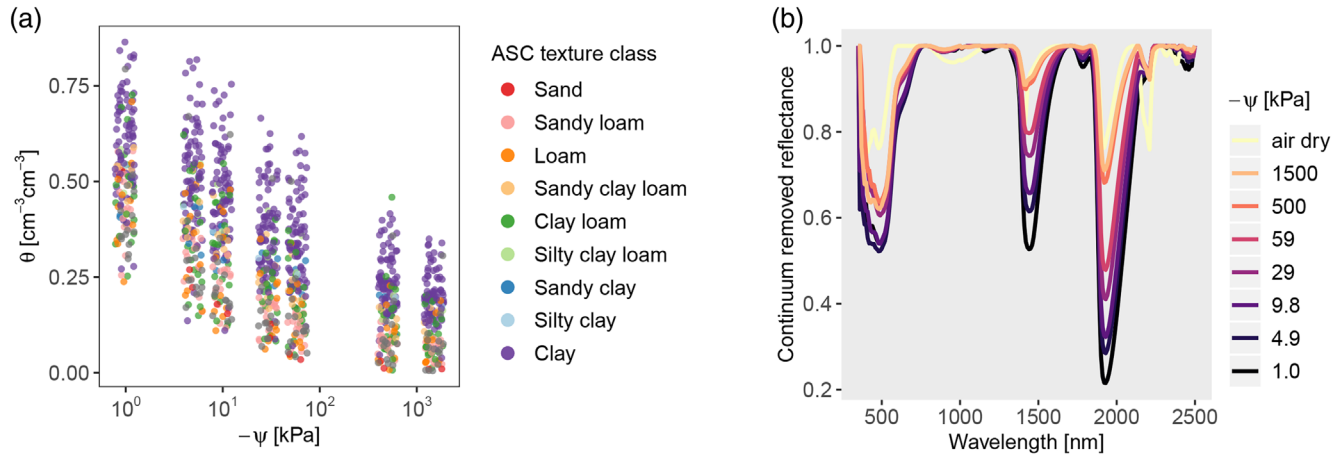


FIGURE 3 (a) Volumetric water contents, θ, at the matric potentials ψ of −1 kPa, −5 kPa, −10 kPa, −29 kPa, −60 kPa, −500 kPa, and −1500 kPa, of the soils representing soil texture classes commonly used in the Australian Soil Classification (ASC) system at the 0–15 cm, 15–30 cm and 30–60 cm depths. (b) Mean continuum removed reflectance spectra by matric potential. The colours of the legend represent the mean volumetric water contents θ for soils at different matric potentials ψ

residual sum of squares, provide the same assessment. To evaluate the final best-fit, we then used the RMSE.

2.4 | Modelling with CUBIST

To derive the vis–NIR models and PTF-LOC, we used the rule-based machine learning algorithm CUBIST (Quinlan, 1992). CUBIST can model complex non-linear relationships between collinear predictor variables (e.g., spectra) and the outcome (e.g., soil water). CUBIST is a tree-derived data partitioning algorithm that models groups of data with piecewise multiple linear regressions. A rule is composed of conditions and linear equations, expressed as if conditions, then linear formula. CUBIST simplifies rules by pruning to remove or merge parts of rules to improve performance further and to prevent over-fitting. Different trees can be built sequentially, where the outcome for the next tree is adjusted depending on the results of the previous tree. These sets of trees, also called committees, can be aggregated by model averaging to reduce variance in the prediction. Further, it is possible to adjust the prediction of new cases with several nearest neighbouring samples in the training set (Quinlan, 1992). For a description of CUBIST in spectroscopic modelling see Viscarra Rossel and Webster (2012). To gain insight into the spectroscopic modelling of the Kosugi SWR parameters, we used the mean of the variable usage statistics for variables in the CUBIST conditions and linear regressions.

2.4.1 | Model tuning and validation

We tuned the CUBIST models using a full-factorial combination of 5, 10 and 20 committees, and 2, 5, 7, and

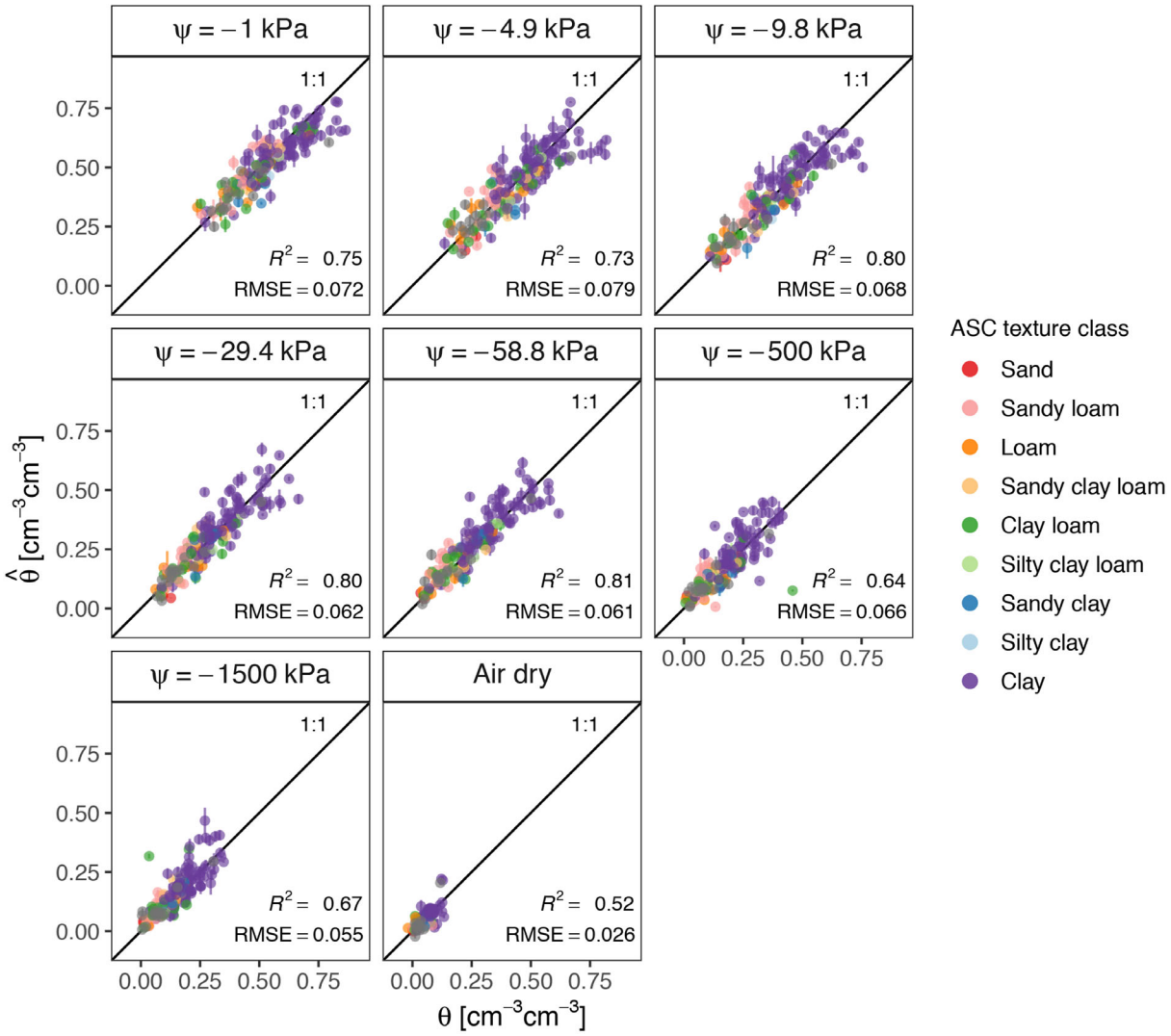


FIGURE 4 Estimates of volumetric water contents, θ, derived with CUBIST, coloured by Australian Soil Classification (ASC) texture classes. The cross-validation root-mean-square-error (RMSE) and R^2 of the model were 0.064 $cm^3\ cm^{-3}$ and 0.89, respectively. Estimates are plotted by matric potential, ψ. The error bars signify prediction intervals provided as standard errors from cross-validated estimates

9 neighbours. We used the CUBIST implementation in the Cubist R package (Kuhn & Quinlan, 2018). Model tuning was performed by minimising the RMSE. The validation of all of the models (including PTF-LOC) involved two nested resampling procedures. We used three repeats of nested 10-fold cross-validation with ten bootstraps on each holdout dataset to separate tuning and final model assessment (Stone, 1974; Varma & Simon, 2006). The hold-out data in each cross-validation iteration (10%) were used as independent subsets to measure model performance after tuning and fitting the models on bootstraps of the corresponding sets (90%). To prevent overly optimistic assessment that might result from potential data leakage (across soil layers), the 10-fold cross-validation was constrained by the site. This ensured all depth measurements at a site be included in either the model fitting or in the assessment, not both. This nested and grouped cross-validation scheme ensures unbiased estimates of performance. For clarity, we provide details and a schematic representation of the nested resampling validation approach in Data S1, Figure S1.

Validation of approach A: SPC-AVG (Figure 2) also involved using only the average spectra of measurements at $\psi = -10$ kPa and $\psi = -1500$ kPa (i.e., not only the averaged spectra of soils at all seven matric potentials). We only show the results from the validation with the two measurements as they were insignificantly different from those that used the average of all seven matric potentials. The reason for validating with only the average spectra from these two measurements is that it enables a more practical application of the SPC-AVG approach. In this case, to estimate AWC one

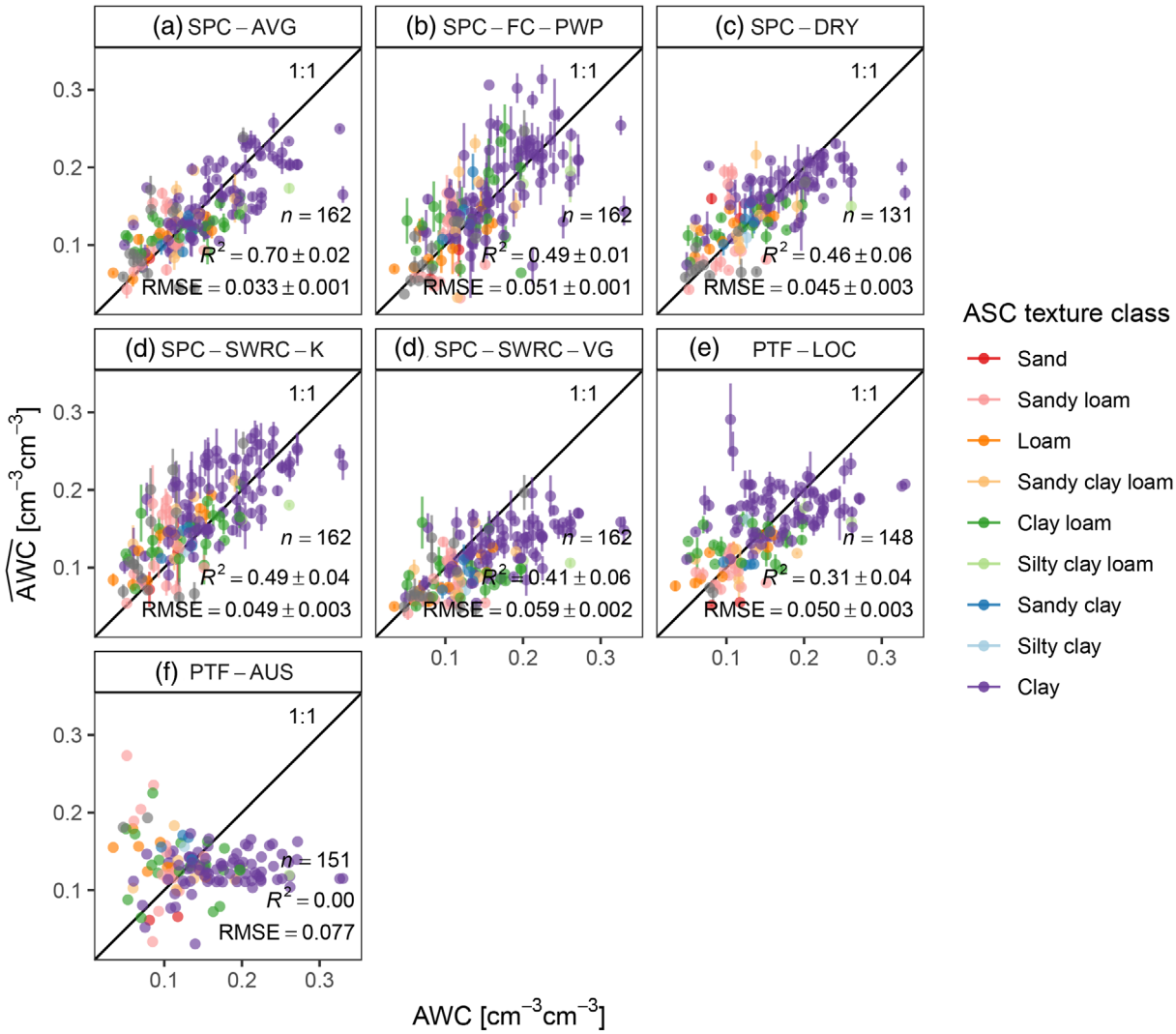


FIGURE 5 Estimates of available water capacity (AWC) with the different approaches: (a) direct spectroscopic modelling of AWC, (b) first modelling water contents at permanent wilting point (θ_{PWP}; $\psi = -1500$ kPa) and at field capacity (θ_{FC}; $\psi = -29.4$ kPa), then calculating AWC (indirectly), (c) spectroscopic modelling of AWC using air-dry spectra, (d) spectroscopic modelling of the parameters of the Kosugi (K) and van Genuchten (VG) water retention models and estimating AWC from the reconstructed curve, (e) estimating AWC with a local pedotransfer function (PTF) using data of this study, and (f) estimating AWC with an established Australian function—See Figure 2. The error bars show standard deviations of the cross-validated predictions. Approach (E: PTF-LOC) has fewer data points because of missing input data (see also Table 1), and approach (C: SPC-DRY) has 31 missing data because dry spectra were not measured

only needs to measure spectra from soils at field conditions.

The statistics used to assess the performance of the machine learning vis–NIR models and the PTF were the RMSE to quantify the inaccuracy of the estimates, the standard deviation of the error (SDE) to quantify their imprecision, and the mean error (ME) to quantify the bias. The RMSE accounts for both the bias and the imprecision of the analysis so that $RMSE^2 = ME^2 + SDE^2$. We also report the coefficient of determination (R^2) from linear regression. We report the mean and standard deviation of the assessment statistics from the cross-validations.

3 | RESULTS

3.1 | Measured soil water contents and spectra

The samples represent a typical range of soil types from Australian agricultural regions (Figure 1), with widely varying physical and chemical properties (Table 1). The measured AWC of the samples ranged from 0.034 to 0.330 $cm^3\ cm^{-3}$ (Table 1).

The volumetric water contents at each matric potential also varied widely, portraying the diverse soil types and textures (Figure 3a; Table 1). At $\psi = -1$ kPa, soil

TABLE 2 Evaluation of the different approaches used to estimate available water capacity (AWC).

Method	Mean	RMSE	ME	SDE	R^2
A: SPC-AVG	0.145	0.033 ± 0.001	−0.001 ± 0.001	0.033 ± 0.001	0.70 ± 0.02
B: SPC-FC-PWP	0.145	0.051 ± 0.001	0.000 ± 0.004	0.051 ± 0.001	0.49 ± 0.01
θ_{PWP}	0.147	0.039 ± 0.003	−0.000 ± 0.001	0.039 ± 0.003	0.78 ± 0.03
θ_{FC}	0.291	0.062 ± 0.001	0.000 ± 0.003	0.062 ± 0.001	0.79 ± 0.01
C: SPC-DRY	0.145	0.045 ± 0.003	−0.000 ± 0.002	0.044 ± 0.003	0.46 ± 0.06
D: SPC-SWRC-K	0.145	0.049 ± 0.003	−0.019 ± 0.002	0.045 ± 0.003	0.49 ± 0.04
D: SPC-SWRC-VG	0.145	0.059 ± 0.002	0.038 ± 0.001	0.046 ± 0.002	0.41 ± 0.06
E: PTF-LOC	0.149	0.050 ± 0.003	0.002 ± 0.000	0.050 ± 0.003	0.31 ± 0.04
F: PTF-AUS	0.148	0.077	0.023	0.074	0.00

Note: SPC-AVG: direct spectroscopic modelling of AWC. SPC-FC-PWP: First modelling water contents at permanent wilting point and at field capacity, then calculating AWC. SPC-DRY: Modelling of AWC with air-dry spectra. SPC-SWRC: Spectroscopic modelling of the parameters of the Kosugi (K) and van Genuchten (VG) models (see Table 3) and estimating AWC from the reconstructed curves. PTF-LOC: Estimating AWC with a local pedotransfer function. PTF-AUS: Estimating AWC with a general PTF derived for Australian soils.

TABLE 3 Evaluation of the vis–NIR soil water retention (SWR) models' parameters

SWR model	Mean	RMSE	ME	SDE	R^2
Kosugi					
θ_s [cm^3cm^{-3}]	0.59	0.082 ± 0.001	−0.012 ± 0.001	0.081 ± 0.001	0.73 ± 0.00
θ_r [cm^3cm^{-3}]	0.13	0.038 ± 0.000	−0.003 ± 0.002	0.038 ± 0.000	0.75 ± 0.01
σ	2.17	0.360 ± 0.004	−0.028 ± 0.006	0.358 ± 0.004	0.39 ± 0.01
ψ_0 [kPa]	−17.2	12.9 ± 0.2	0.1 ± 0.5	12.9 ± 0.2	0.18 ± 0.02
van Genuchten					
θ_s [cm^3cm^{-3}]	0.57	0.076 ± 0.004	−0.010 ± 0.004	0.075 ± 0.003	0.74 ± 0.02
θ_r [cm^3cm^{-3}]	0.14	0.043 ± 0.001	−0.001 ± 0.002	0.043 ± 0.001	0.71 ± 0.01
α [cm^{-1}]	0.04	0.022 ± 0.000	−0.001 ± 0.000	0.022 ± 0.000	0.08 ± 0.01
n	1.56	0.17 ± 0.00	0.01 ± 0.00	0.17 ± 0.00	0.24 ± 0.01

water contents ranged from 0.238 to 0.865 cm^3 cm^{-3}. At wilting point ($\psi = -1500$ kPa), water contents ranged from 0.005 to 0.350 cm^3 cm^{-3}. The continuum removed spectra show that a reduction in soil water content with increasing matric potentials causes a proportional increase in reflectance and narrower absorption features (Figure 3b). Water affects the spectra most at the water absorption wavelengths around 1400 nm and 1900 nm in the NIR region and broadly across the visible range.

3.2 | Prediction of water contents at different matric potentials and available water capacity

The spectroscopic predictions of θ at each of the selected matric potentials (Figure 4), resulted in RMSE values of between 0.026 and 0.081 cm^3 cm^{-3}, and R^2 values from 0.50 to 0.80. The spectroscopic predictions of air-dry water content were relatively accurate but accounted only for a small proportion of measured variance due to its small range (0.000–0.133 cm^3 cm^{-3}).

Direct spectroscopic modelling of AWC (A: SPC-AVG) produced the most accurate estimates, with the smallest RMSE and largest R^2 (RMSE = 0.033 cm^3 cm^{-3}; R^2 = 0.70; Figure 5; Table 2). First modelling water contents at FC and PWP and then calculating AWC (B: SPC-FC-PWP) produced the least accurate spectroscopic estimates, with an RMSE of 0.051 cm^3 cm^{-3} and R^2 of 0.49. Individual spectroscopic estimates of PWP and FC were relatively accurate with R^2 values of 0.78 and 0.79, respectively (Figure 5; Table 2). Estimates of AWC with the air-dry spectra (C: SPC-DRY) produced an RMSE of 0.045 cm^3 cm^{-3} and R^2 of 0.46. Estimates of AWC following spectroscopic modelling of the parameters of the Kosugi SWR model and then reconstructing the water retention curve to estimate AWC (D: SPC-SWRC-K) had an RMSE of 0.049 cm^3 cm^{-3} and R^2 of 0.49. The RMSE of

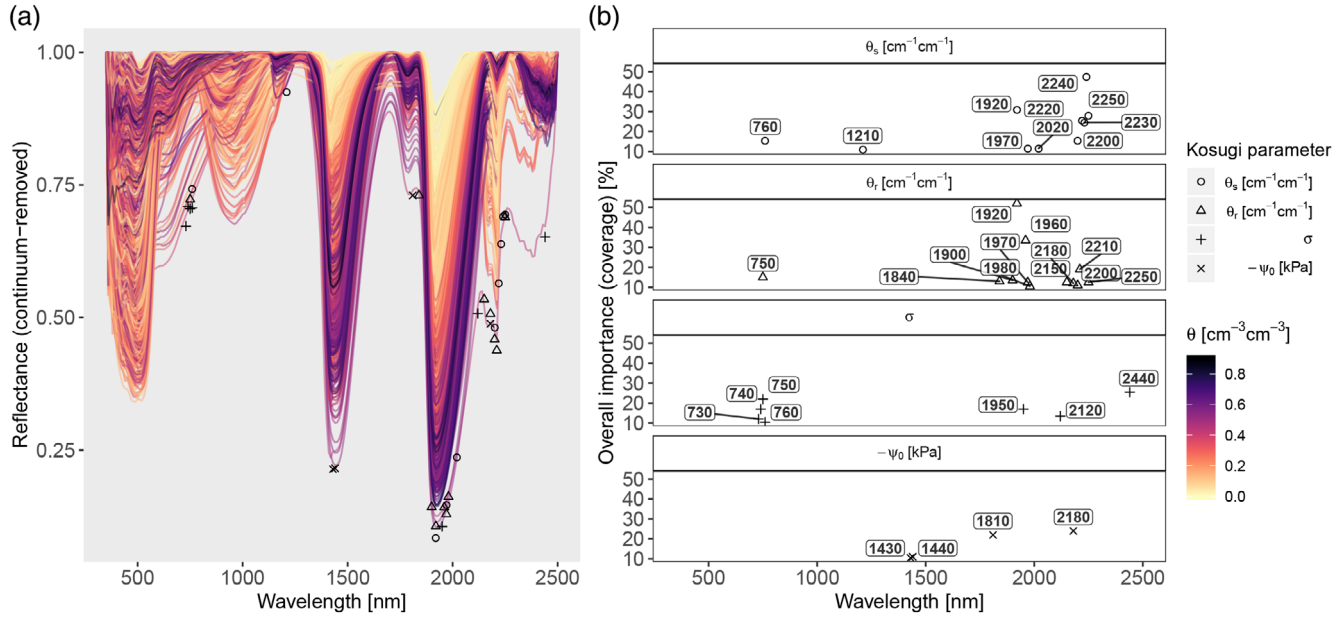


FIGURE 6 Interpretation of vis–NIR modelled Kosugi parameters θ_s, θ_r, σ, and ψ_0. Continuum removed reflectance spectra (a), coloured by volumetric water content θ, are shown with important variables for Kosugi parameters (b). Influential spectral variables were determined with CUBIST overall importance (percent coverage in split conditions and regressions) higher than 10%

the AWC estimates from the reconstruction of the van Genuchten model was 0.059 cm^3 cm^{-3} and the R^2 was 0.41 (Figure 5; Table 2). The spectroscopic models were generally more imprecise than biased, although the estimates of AWC from the reconstructed SWR curves were somewhat more biased (Table 2).

Estimates of AWC with the local PTF (E: PTF-LOC) had an RMSE of 0.050 cm^3 cm^{-3} and R^2 of 0.31, while those from the general PTF (F: PTF-AUS) were the least accurate with an RMSE of 0.077 cm^3 cm^{-3} and R^2 of ca. zero as it failed to account for the variation in AWC across the sites (Figure 5; Table 2).

3.3 | vis–NIR estimates of the soil water retention model parameters and reconstruction

For both the Kosugi and van Genuchten models, saturated and residual water contents were well explained by the CUBIST models ($R^2 = 0.71$–0.75; Table 3). Estimates of Kosugi σ were relatively unbiased but the model explained only 39% of its variability. Estimates of the inflection point ψ_0 were more biased and the R^2 of the predictions was 0.18 ± 0.02. The spectroscopic model explained only 8% of the variability in the van Genuchten α parameter, and estimates of the pore-size distribution parameter, n, were poor ($R^2 = 0.28$; Table 3).

The SWR curves for all of the samples, fitted with both the Kosugi and van Genuchten models, as well as their vis–NIR reconstructions, were used to estimate AWC as per approach D: SPC-SWRC are given in the Data S1 (Figures S2, S3, S4, S5).

Most of the important wavelengths (overall relative importance greater than 10%) for the spectroscopic estimates of the Kosugi parameters were centred closely around the peak maxima of the vis–NIR water absorption wavelengths (1200, 1400, 1900, 2200 nm), as well as in-between peak edge features (e.g., 750, 760, 2020 nm) (Figure 6).

4 | DISCUSSION

4.1 | vis–NIR spectroscopy and soil water retention

Diffuse reflectance spectra in the vis–NIR region responds to changes in soil water (Baumgardner et al., 1986; Bowers & Hanks, 1965; Bowers & Smith, 1972; Peterson & Baumgardner, 1981) and it is possible to model gravimetric soil water content with vis–NIR spectra (Lobell & Asner, 2002). The diffuse reflectance of soil around the main water absorption regions (1440 and 1920 nm) decreases non-linearly with increasing water content (Figure 5). This is due to matrix effects that result from the simultaneous presence and physicochemical interaction of soil organic and mineral constituents and water (Lobell & Asner, 2002; Weidong et al., 2002; Whiting et al., 2004). The response may also involve physical scattering phenomena that are influenced

by particle size distribution, surface geometry, pore spacing, the form of water (free, in mineral lattices, surface-bound) and hydrophobicity, due to the presence of organic matter (Knadel et al., 2014).

More recently, studies have also reported the potential to model AWC and SWR with spectra (Babaeian, Homaee, Montzka, et al., 2015; Pittaki-Chrysodonta et al., 2018; Santra et al., 2009). However, there is little published research on modelling of AWC and SWR characteristics with machine learning over a large geographic extent and using a diverse range of soil types with highly variable chemical composition and textures. This research uses samples from 10 of the 14 orders in the Australian soil classification system, which are the most important for agricultural production (Figure 1). We described experiments that showed that their water contents, measured at seven different potentials from $\psi = -1$ kPa to $\psi = -1500$ kPa and under air-dried conditions, can be well estimated with vis–NIR spectroscopy and machine learning. Our experiments also provide a comprehensive assessment of different approaches for estimating AWC. This includes the spectral reconstruction of the SWR curve and comparisons to estimates based on local and general PTFs.

4.2 | Spectroscopic estimates of soil water content

Our estimates of θ were unbiased, but somewhat less accurate when the soils were drier ($\psi = -58.8$ to -1500 kPa; Figure 4, Table 2), because of the stronger water absorption signals of the wetter soils (primarily 1440 nm, 1920 nm), compared to when they were drier ($\psi = -500$ to -1500 kPa). In these cases, wavelengths related to water were less prominent, and the estimates relied more on indirect relationships to wavelengths that correspond to their mineral-organic composition (Stenberg et al., 2010; Viscarra Rossel et al., 2009).

Direct modelling of AWC with the spectra (SPC-AVG) produced the most accurate estimates (RMSE = 0.033 $cm^3\ cm^{-3}$). The reason might be that the models use mainly direct relationships to the most distinct spectral features of water (see Figure 3). Estimates of AWC from vis–NIR models of θ_{FC} and θ_{PWP} (SPC-FC-PWP) were the least accurate of the spectroscopic approaches (RMSE = 0.051 $cm^3\ cm^{-3}$). The likely reason is that fitting two models (for FC and PWP) adds more uncertainty to the final estimates. The error of the model that used air-dry spectra (SPC-DRY) to estimate AWC (RMSE = 0.045; $R^2 = 0.46$; measured = 0.034–0.036 $cm^3\ cm^{-3}$) tended to underestimate larger AWCs (Figure 4). The reason might be that the model built with air-dry spectra relied more on indirect relationships to the mineral-organic composition of the soils, which were not strong enough to model larger AWCs. Nevertheless, the RMSE of our estimates were similar to other published studies (e.g., Blaschek et al., 2019), although as mentioned above, our study is with a more diverse set of soils.

4.3 | Prediction of water contents at different matric potentials and available water capacity

Estimates of AWC from vis–NIR models of the Kosugi model parameters (SPC-SWRC-K) were less accurate than the SPC-AVG and only slightly less accurate than the SPC-DRY approach (Figure 4). Estimates of AWC from the reconstruction of the van Genuchten model were less accurate than those from the Kosugi model. This might be because most soils lacked a distinct air-entry value (see Figure S4) and also because there was no relationship to soil properties that can be modelled well with the vis–NIR spectra. The spectroscopic modelling of n was poorer compared to the spectroscopic modelling of its Kosugi counterpart σ. The reason for this difference might be that the Kosugi model incorporates a log-normal pore size distribution function, which could have generalised SWR better than using the van Genuchten, which has the empirical fitting parameter n. Parameter n was better predicted with air-dry spectra in Santra et al. (2009), however, the soil samples covered a regional study area with a defined soil type, and hence results cannot directly be compared to the variability of our data set.

Compared to the other approaches, the SPC-SWRC approach is more versatile because it enables the derivation of the entire SWR curve and rapid predictions of water content at specific matric potentials. The SPC-SWRC approach can be more useful in applications that also require data on water contents at specific matric potentials, e.g., biogeochemical and hydrological modelling, the assessment of soil conditions using proximal sensing platforms (Viscarra Rossel et al., 2017), continuously integrating water contents in lysimeter and soil column experiments across depth, digital soil mapping, or to infer hydraulic conductivity (Assouline & Or, 2013; Szabó et al., 2018; Vereecken et al., 2016). The PTF estimates were the least accurate. Estimates with the PTF-LOC were imprecise and biased, but better than those of the PTF-AUS. The reason for the poorer predictability of the PTF-LOC, compared to the spectroscopic methods, might be that the information content of the input soil properties is smaller than that of the spectra. Thus, the predictor space of the soil properties would limit specific predictive relationships with soil water, leading to difficulties in representing local conditions. For similar reasons, Babaeian, Homaee, Vereecken, et al. (2015) also reported

better predictions of water contents and SWR with vis–NIR than with PTFs.

4.4 | Spectral interference systems and direct spectroscopic modelling

McBratney et al. (2006) proposed the development of spectral inference systems to infer soil properties that are difficult to measure, such as SWR and AWC. These systems use infrared spectra to derive soil properties for input into PTFs (e.g., Tranter et al., 2008). Depending on the soil properties to be inferred, the approach will be more or less useful. For predictions of SWR and AWC, it is likely that the propagated uncertainties of the estimates from such spectral PTF engines will be larger compared to direct spectroscopic modelling, limiting the usefulness of the estimates. Modelling the spectra with CUBIST directly to predict SWR and AWC, like we have done here, removes the intermediate step and result in smaller uncertainties and more useful data. CUBIST models were able to identify and use wavelengths that are relevant to soil water and SWR across water potentials. They allowed embedding of both causal and indirect non-linear interactions between the spectra, which represents the soil's mineral-organic composition, water and particle size, and soil water and its characteristics, across a diverse range of soil types.

4.5 | Modelling with CUBIST

The spectroscopic models could predict the saturated θ_s ($R^2 = 0.73$ and 0.74; RMSE = 0.082 and 0.076 $cm^3\ cm^{-3}$) and residual θ_r ($R^2 = 0.75$ and 0.71; RMSE = 0.038 and 0.043 $cm^3\ cm^{-3}$) water content parameters of the Kosugi and van Genuchten models (Table 3). The θ_s parameter represents the effective porosity and is defined by capillary forces, while θ_r mostly depends on adsorptive processes, which are related to texture and the specific surface area of soil (Tuller & Or, 2005). The estimates were accurate because the models could use direct relationships to the water absorption wavelengths (e.g. near 1900 nm; Figure 6), and relationships to particle size distribution and mineralogy (e.g., those near 2200 nm; Figure 6), which determine the soil's specific surface area (Knadel et al., 2014, 2020).

Babaeian, Homaee, Montzka, et al. (2015) used the vis–NIR spectra from dry soil to model van Genuchten θ_s within a region with similar soil types and found poor predictability (validation $R^2 = 0.1$; RMSE = 0.062 $cm^3\ cm^{-3}$), given a measured range of 0.036–0.061 $cm^3\ cm^{-3}$ (standard deviation = 0.062 $cm^3\ cm^{-3}$). Pittaki-Chrysodonta et al. (2018) reported better predictability of θ between θ_s and θ_r ($\Psi = -98$ kPa; RMSE = 0.022 $cm^3\ cm^{-3}$; $R^2 = 0.93$), however, the experiment and assessment was limited to within $\Psi = -1$ kPa and -98 kPa and to soil samples from a narrow range of soils from six agricultural sites.

Estimates of σ were unbiased but imprecise ($R^2 = 0.39$) because the models could only rely on a few, indirect relationships to water and clay mineralogy (Figure 6). The Kosugi σ parameter, which is related to the width of the log-normal pore radius distribution (Kosugi, 1994), may be characterised by the fractional content of the particle size distribution, soil structure and their interactions. The attribution to the conceptual pore size distribution made for Kosugi σ shows similar effects as the empirical shape or slope parameter b of the Campbell model, for which Pittaki-Chrysodonta et al. (2018) achieved good results ($R^2 = 0.90$ and 0.89) with vis–NIR and pedotransfer models. Estimates of the ψ_0 parameter, which represents the capillary pressure at the inflection point, were inaccurate ($R^2 = 0.18$) because the models could not find strong relationships to information in the spectra (Figure 6)—spectroscopy is essentially a surface technique Norouzi et al. (2021). The spectrally reconstructed SWR curves at the measured matric potentials showed no overall trend in the error distributions and range between the soil texture classes (Figure 6).

4.6 | Using repacked core samples

Repacked soil core samples do not possess macropores, which are important because they affect capillary forces, which can enable the soil to potentially hold more water at a given suction. Repacking also alters micro- and macro-aggregates in soil and the pore space between aggregates. Hence, measurements of SWR using intact soil cores are different from measurements using repacked cores, particularly at lower matric potentials. Micropore size distributions and continuity, however, may be less affected by repacking in many soils because the radii of smaller particles do not change as much, particularly if the repacking is done to bulk densities that approximate those at field conditions (Flint & Flint, 2002).

Our experiments used repacked soil samples, which do not precisely reproduce field conditions and structure. Measurements of the wet end of the SWR were more affected than those of the dry end, which were well reconstructed by the spectroscopic method (Tables 1 and 2). The reason for using repacked cores was that the experiments involved a diverse set of soil types from profiles sampled over a large geographical extent, which extends from southern Queensland to New South Wales, South Australia and Western

Australia (Figure 1). In such cases, the use of repacked cores for measuring SWR is not uncommon as they provide a scientifically valid and practical procedure, for approximating hydraulic properties Gupta and Larson (1979). Additionally, measurements with pressure plates are more practical using repacked cores because of the difficulties encountered with using intact cores (Dane & Hopmans, 2002; Gupta & Larson, 1979).

In our experimental setup with repacked cores, a number of the high clay content Vertosols samples had exceptionally high volumetric water contents (>0.65 cm^3 cm^{-3}). The reason is that these soils exhibit shrink-swell behaviour, but we could not quantify the change in soil volume and the measured bulk densities were unrepresentative of soil under field conditions. However, this shortcoming of our experiment does not affect our research to explore the relationship between SWR and spectra and investigate different approaches for the spectroscopic modelling of AWC and SWR.

Spectroscopic methods measure only the soil surface, therefore they cannot "see" pores or structure. If we were to use a spectroscopic method in the field, one would measure only the soil surface, be it top or subsoil, as measurements down the profile in an open soil pit, or the surface of sampled soil cores (Viscarra Rossel et al., 2017).

4.7 | Spectroscopic approaches have the potential for estimating soil water retention and water-holding capacity

Our findings are significant for two main reasons. The presented methods can help to meet the enormous demand for data on soil water at different scales and different applications. With appropriate calibrations, the spectroscopic-machine learning method can also help to overcome the substantial cost and complexity of conventional laboratory and field measurements of SWR and AWC. The methods could also be used directly on soil under field conditions by proximal sensing; in situ or ex-situ (Viscarra Rossel et al., 2017).

5 | CONCLUSION

The machine learning models were able to capture the direct and indirect relationships between the vis–NIR spectra and θ, AWC and SWR, of a diverse set of Australian agricultural soils. Direct modelling of AWC with vis–NIR spectra produced the most accurate estimates with the smallest RMSE. Calculation of AWC from the estimates of spectroscopic models of FC and PWP was less accurate because of the errors of the two separate spectroscopic models. Estimates of AWC with air-dry spectra tended to slightly underestimate samples with more AWC. Estimates of AWC using reconstructed SWR curves from spectroscopic estimates of the parameters of SWR functions were less accurate than the direct estimates of AWC. However, the approach is more versatile, particularly if one necessitates data on water content at different matric potentials, or if one needs to describe the soil SWR characteristic. Estimates from the general PTFs were the least accurate. The spectroscopic methods that we present can help with the cost-effective collection of information on SWR and AWC and future research might aim to operationalise one of the approaches for in-field proximal sensing.

ACKNOWLEDGEMENTS

We thank Dr Yakov A. Pachepsky for his insightful comments on an early draft of our manuscript. Open access publishing facilitated by Curtin University, as part of the Wiley - Curtin University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Philipp Baumann: Formal analysis (equal); methodology (supporting); validation (equal); writing – original draft (equal); writing – review and editing (supporting). **Juhwan Lee:** Formal analysis (supporting); investigation (supporting); writing – review and editing (supporting). **Thorsten Behrens:** Formal analysis (supporting); investigation (supporting); validation (supporting); writing – review and editing (supporting). **Asim Biswas:** Formal analysis (supporting); validation (supporting); writing – review and editing (supporting). **Johan Six:** Formal analysis (supporting); investigation (supporting); writing – review and editing (supporting). **Gordon Mclachlan:** Data curation (lead). **Raphael A. Viscarra Rossel:** Conceptualization (lead); methodology (lead); formal analysis (equal); writing – original draft (equal); writing – review and editing (lead).

DATA AVAILABILITY STATEMENT

Data is available from corresponding author upon reasonable request.

ORCID

Philipp Baumann https://orcid.org/0000-0002-3194-8975
Juhwan Lee https://orcid.org/0000-0002-7967-2955
Raphael A. Viscarra Rossel https://orcid.org/0000-0003-1540-4748

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (pp. 267–281). Springer.

Assouline, S., & Or, D. (2013). Conceptual and parametric representation of soil hydraulic properties: A review. *Vadose Zone Journal*, *12*(4), 1–20.

Babaeian, E., Homaee, M., Montzka, C., Vereecken, H., & Norouzi, A. A. (2015). Towards retrieving soil hydraulic properties by hyperspectral remote sensing. *Vadose Zone Journal*, *14*(3), 1–17.

Babaeian, E., Homaee, M., Vereecken, H., Montzka, C., Norouzi, A. A., & van Genuchten, M. T. (2015). A comparative study of multiple approaches for predicting the soil—Water retention curve: Hyperspectral information vs. basic soil properties. *Soil Science Society of America Journal*, *79*(4), 1043–1058.

Baumgardner, M. F., Silva, L. F., Biehl, L. L., & Stoner, E. R. (1986). Reflectance properties of soils. In N. C. Brady (Ed.), *Advances in agronomy* (Vol. 38, pp. 1–44). Academic Press.

Ben-Dor, E., Irons, J. R., & Epema, G. F. (1999). Soil reflectance. In N. Rencz (Ed.), *Soil reflectance. Vol. 3 of remote sensing for the earth sciences: Manual of remote sensing*. John Wiley & Sons.

Bishop, J. L. (1994). Infrared spectroscopic analyses on the nature of water in montmorillonite. *Clays and Clay Minerals*, *42*(6), 702–716.

Blaschek, M., Roudier, P., Poggio, M., & Hedley, C. B. (2019). Prediction of soil available water-holding capacity from visible near-infrared reflectance spectra. *Scientific Reports*, *9*(1), 12833.

Bouma, J., 1989. Using soil survey data for quantitative land evaluation. In: Stewart, B. A. (Ed.), *Advances in soil science*. Advances in Soil Science. Springer US, pp. 177–213.

Bowers, S. A., & Hanks, R. J. (1965). Reflection of radiant energy from soils. *Soil Science*, *100*(2), 130–138.

Bowers, S. A., & Smith, S. J. (1972). Spectrophotometric determination of soil water content. *Soil Science Society of America Journal*, *36*(6), 978–980.

Bowman, G., & Hutka, J. (2002). Particle size analysis. In N. McKenzie, K. Coughlan, & H. Cresswell (Eds.), *Soil physical measurement and interpretation for land evaluation* (pp. 224–239). CSIRO Publishing.

Clark, R. N., & Roush, T. L. (1984). Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research: Solid Earth*, *89*(B7), 6329–6340.

Curcio, J. A., & Petty, C. C. (1951). The near infrared absorption spectrum of liquid water. *JOSA*, *41*(5), 302–304.

Dane, J. H., & Hopmans, J. W. (2002). 3.3.2 laboratory. In *Methods of soil analysis* (pp. 675–720). John Wiley & Sons, Ltd.

Flint, L. E., & Flint, A. L. (2002). 2.3 Porosity. In *Methods of soil analysis* (pp. 241–254). John Wiley & Sons, Ltd.

Gee, G. W., & Bauder, J. W. (1986). Particle size analysis. In A. Klute (Ed.), *Methods of soil analysis. Part 1* (2nd ed.. Agronomy Monograph no. 9, pp. 383–411). American Society of Agronomy and Soil Science Society of America, Madison.

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

Isbell, R., 2002. *The Australian Soil Classification (revised edition)*. Melbourne: CSIRO Publishing.

Knadel, M., de Jonge, L. W., Tuller, M., Rehman, H. U., Jensen, P. W., Moldrup, P., Greve, M. H., & Arthur, E. (2020). Combining visible near-infrared spectroscopy and water vapor sorption for soil specific surface area estimation. *Vadose Zone Journal*, *19*(1), e20007.

Knadel, M., Deng, F., Alinejadian, A., Wollesen de Jonge, L., Moldrup, P., & Greve, M. H. (2014). The effects of moisture conditions—From wet to hyper dry—On visible near-infrared spectra of danish reference soils. *Soil Science Society of America Journal*, *78*(2), 422–433.

Kosugi, K. (1994). Three-parameter lognormal distribution model for soil water retention. *Water Resources Research*, *30*(4), 891–901.

Kuhn, M., & Quinlan, R., 2018. Cubist: Rule- and instance-based regression modeling. R package version 0.2.2. URL https://CRAN.R-project.org/package=Cubist

Lobell, D. B., & Asner, G. P. (2002). Moisture effects on soil reflectance. *Soil Science Society of America Journal*, *66*, 6–727.

Manzoni, S., Schimel, J. P., & Porporato, A. (2012). Responses of soil microbial communities to water stress: Results from a meta-analysis. *Ecology*, *93*(4), 930–938. https://doi.org/10.1890/11-0026.1

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441.

McBratney, A. B., Minasny, B., Viscarra Rossel, R., Dec. 2006. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma 136* (1), 272–278.

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*(3–4), 225–253.

Norouzi, S., Sadeghi, M., Liaghat, A., Tuller, M., Jones, S. B., & Ebrahimian, H. (2021). Information depth of NIR/SWIR soil reflectance spectroscopy. *Remote Sensing of Environment*, *256*, 112315.

Pachepsky, Y., & Hill, R. L. (2017). Scale and scaling in soils. *Geoderma*, *287*, 4–30.

Pachepsky, Y., & Rawls, W. J. (2004). *Development of Pedotransfer functions in soil hydrology* (Vol. 30). Elsevier.

Padfield, D., Matheson, G., 2020. nls.multstart: Robust Non-Linear Regression using AIC Scores. R package version 1.2.0. URL https://CRAN.R-project.org/package=nls.multstart

Passioura, J. B. (2002). Soil conditions and plant growth. *Plant, Cell & Environment*, *25*(2), 311–318.

Peterson, J., & Baumgardner, M. (1981). Use of spectral data to estimate the relationship between soil moisture tensions and their corresponding reflectances. *Annual Report OWRT Purdue University (No. 143)*, 1–18.

Pittaki-Chrysodonta, Z., Moldrup, P., Knadel, M., Iversen, B. V., Hermansen, C., Greve, M. H., & de Jonge, L. W. (2018). Predicting the Campbell soil water retention function: Comparing visible-near-infrared spectroscopy with classical pedotransfer function. *Vadose Zone Journal*, *17*(1), 1–12.

Quinlan, J. R. (1992). Learning with continuous classes. In *In: 5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343–348). World Scientific.

Rayment, G. E., & Higginson, F. R. (1992). *Australian laboratory handbook of soil and water chemical methods. Australian soil and land survey handbooks series*. Inkata Press.

Santra, P., Sahoo, R. N., Das, B. S., Samal, R. N., Pattanaik, A. K., & Gupta, V. K. (2009). Estimation of soil hydraulic properties using proximal spectral reflectance in visible, near-infrared, and shortwave-infrared VIS-NIR-SWIR region. *Geoderma*, *152*(3), 338–349.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639.

Shein, E. V., & Arkhangel'skaya, T. A. (2006). Pedotransfer functions: State of the art, problems, and outlooks. *Eurasian Soil Science*, *39*(10), 1089–1099.

Skopp, J., Jawson, M. D., & JW, D. (1990). Steady-state aerobic microbial activity as a function of soil water content. *Soil Science Society of America Journal*, *54*, 1619–1625.

Soltani, I., Fouad, Y., Michot, D., Bréger, P., Dubois, R., & Cudennec, C. (2018). A near infrared index to assess effects of soil texture and organic carbon content on soil water content: Effect of texture and organic carbon on soil water with nir. *European Journal of Soil Science.*, *70*, 151–161.

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. V., Macdonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews*, *49*(2), 139–186.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, *107*. Elsevier, 163–215.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 111–147 URL http://www.jstor.org/stable/2984809

Stoner, E. R., Baumgardner, M. F., Weismiller, R. A., Biehl, L. L., & Robinson, B. F. (1980). Extension of laboratory-measured soil spectra to field conditions. *Soil Science Society of America Journal*, *44*(3), 572–574.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., & Pásztor, L. (2018). Mapping soil hydraulic properties using random Forest based pedotransfer functions and geostatistics. *Hydrology and Earth System Sciences Discussions*, *23*, 2615–2635.

Tranter, G., Minasny, B., McBratney, A. B., Viscarra Rossel, R. A., & Murphy, B. W. (2008). Comparing spectral soil inference systems and mid-infrared spectroscopic predictions of soil moisture retention. *Soil Science Society of America Journal*, *72*(5), 1394–1400.

Tuller, M., & Or, D. (2005). Water films and scaling of soil characteristic curves at low water contents. *Water Resources Research*, *41*(9), W09403.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, *44*(5), 892–898.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., Schaap, M. G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M. J., Weihermüller, L., Zacharias, S., Zhang, Y., & Vereecken, H. (2017). Pedotransfer functions in earth system science: Challenges and perspectives. *Reviews of Geophysics*, *55*(4), 1199–1256.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91.

Vereecken, H., Huisman, J. A., Bogena, H., Vanderborght, J., Vrugt, J. A., & Hopmans, J. W. (2008). On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research*, *44*(4), W00D06.

Vereecken, H., Schnepf, A., Hopmans, J., Javaux, M., Or, D., Roose, T., Vanderborght, J., Young, M., Amelung, W., Aitkenhead, M., Allison, S., Assouline, S., Baveye, P., Berli, M., Brüggemann, N., Finke, P., Flury, M., Gaiser, T., Govers, G., ... Young, I. (2016). Modeling soil processes: Review, key challenges, and new perspectives. *Vadose Zone Journal*, *15*(5), 1–57.

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattе, J. A. M., Shepherd, K. D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B. G., Bartholomeus, H. M., Bayer, A. D., Bernoux, M., Böttcher, K., Brodsky, L., Du, C. W., Chappell, A., ... Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, *155*, 198–230.

Viscarra Rossel, R. A., Cattle, S. R., Ortega, A., & Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy. *Geoderma*, *150*(3), 253–266.

Viscarra Rossel, R. A., Lobsey, C. R., Sharman, C., Flick, P., & McLachlan, G. (2017). Novel proximal sensing for monitoring soil organic C stocks and condition. *Environmental Science & Technology*, *51*(10), 5630–5641.

Viscarra Rossel, R. A., & McBratney, A. B. (1998). Laboratory evaluation of a proximal sensing technique for simultaneous measurement of soil clay and water content. *Geoderma*, *85*(1), 19–39.

Viscarra Rossel, R. A., & Webster, R. (2012). Predicting soil properties from the australian soil visible-near infrared spectroscopic database. *European Journal of Soil Science*, *63*(6), 848–860.

Weidong, L., Baret, F., Xingfa, G., Qingxi, T., Lanfen, Z., & Bing, Z. (2002). Relating soil surface moisture to reflectance. *Remote Sensing of Environment*, *81*(2–3), 238–246.

Whiting, M. L., Li, L., & Ustin, S. L. (2004). Predicting water content using Gaussian model on soil spectra. *Remote Sensing of Environment*, *89*(4), 535–552.

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*, 123–150.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Baumann, P., Lee, J., Behrens, T., Biswas, A., Six, J., McLachlan, G., & Viscarra Rossel, R. A. (2022). Modelling soil water retention and water-holding capacity with visible–near-infrared spectra and machine learning. *European Journal of Soil Science*, *73*(2), e13220. https://doi.org/10.1111/ejss.13220