Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Combining visible near-infrared spectroscopy and water vapor sorption for soil specific surface area estimation

Knadel, Maria; de Jonge, Lis Wollesen; Tuller, Markus; Rehman, Hafeez Ur; Jensen, Peter Weber; Moldrup, Per; Greve, Mogens H.; Arthur, Emmanuel

Link to publication from Aalborg University

ORIGINAL RESEARCH ARTICLE

Combining visible near-infrared spectroscopy and water vapor sorption for soil specific surface area estimation

Maria Knadel[1] | **Lis Wollesen de Jonge[1]** | **Markus Tuller[2]** | **Hafeez Ur Rehman[1]** | **Peter Weber Jensen[1]** | **Per Moldrup[3]** | **Mogens H. Greve[1]** | **Emmanuel Arthur[1]**

[1]Dep. of Agroecology, Aarhus Univ., Blichers Allé 20, DK-8830 Tjele, Denmark

[2]Dep. of Environmental Science, The Univ. of Arizona, 1177 E. 4th St., Tucson, AZ 85721, USA

[3]Dep. of Civil Engineering, Aalborg Univ., Thomas Manns Vej 23, Aalborg Ø, 9200, Denmark

Correspondence
Maria Knadel, Dep. of Agroecology, Aarhus Univ., Blichers Allé 20, DK-8830 Tjele, Denmark
Email: maria.knadel@agro.au.dk

Abstract

The soil specific surface area (SSA) is a fundamental property governing a range of soil processes relevant to engineering, environmental, and agricultural applications. A method for SSA determination based on a combination of visible near-infrared spectroscopy (vis-NIRS) and vapor sorption isotherm measurements was proposed. Two models for water vapor sorption isotherms (WSIs) were used: the Tuller–Or (TO) and the Guggenheim–Anderson–de Boer (GAB) model. They were parameterized with sorption isotherm measurements and applied for SSA estimation for a wide range of soils ($N = 270$) from 27 countries. The generated vis-NIRS models were compared with models where the SSA was determined with the ethylene glycol monoethyl ether (EGME) method. Different regression techniques were tested and included partial least squares (PLS), support vector machines (SVM), and artificial neural networks (ANN). The effect of dataset subdivision based on EGME values on model performance was also tested. Successful calibration models for SSA_{TO} and SSA_{GAB} were generated and were nearly identical to that of SSA_{EGME}. The performance of models was dependent on the range and variation in SSA values. However, the comparison using selected validation samples indicated no significant differences in the estimated SSA_{TO}, SSA_{GAB}, and SSA_{EGME}, with an average standardized RMSE (SRMSE = RMSE/range) of 0.07, 0.06 and 0.07, respectively. Small differences among the regression techniques were found, yet SVM performed best. The results of this study indicate that the combination of vis-NIRS with the WSI as a reference technique for vis-NIRS models provides SSA estimations akin to the EGME method.

1 | INTRODUCTION

The soil specific surface area (SSA) plays a crucial role for a wide range of soil processes, including the movement and retention of water, nutrient, and contaminant dynamics, ion exchange reactions, microbial activity, heat transport, development of soil structure, and geotechnical soil behavior (Pennell, 2002; Petersen, Moldrup, Jacobsen, & Rolston,

Abbreviations: ANN, artificial neural network(s); BET, Brunauer–Emmet–Teller; EGME, ethylene glycol monoethyl ether; GAB, Guggenheim–Anderson–de Boer; PLS, partial least square(s); SOC, soil organic carbon; SRMSE, standardized root mean square error; SSA, soil specific surface area; SVM, support vector machine(s); TO, Tuller–Or; vis-NIRS, visible near-infrared spectroscopy; VSA, vapor sorption analyzer; WSI, water vapor sorption isotherm.

1996). The SSA is expressed as surface area per unit mass of soil ($m^2\ g^{-1}$). Depending on the organic and mineral composition and particle size distribution of the soil, the values of SSA can differ greatly (Pennell, 2002). In general, soils with elevated clay contents exhibit large SSA, whereas sandy soils have much smaller SSA (Petersen et al., 1996). Moreover, for a given sample, the measurement technique itself can affect the estimates of SSA. The techniques to measure SSA include both direct and indirect methods. Direct estimations are performed by measuring the size and shape of soil particles (Borkovec, Wu, Degovics, Laggner, & Sticher, 1993). Indirect techniques comprise gas-phase adsorption (N_2, CO_2, C_2H_6, C_2H_4, and C_2H_2) (de Jonge & Mittelmeijer-Hazeleger, 1996; de Jonge, de Jonge, & Mittelmeijer-Hazeleger, 2000; Kim, Yoon, & Bae, 2016) and retention of polar liquids such as water (Amali, Petersen, & Rolston, 1994; Arthur et al., 2018; Tuller & Or, 2005), ethylene glycol, ethylene glycol monoethyl ether (EGME) (Cerato & Lutenegger, 2002; Knadel et al., 2018; Petersen et al., 1996), and methylene blue (Hang & Brindley, 1970), with the EGME method being the most common (Pennell, 2002). Apart from water, the use of other polar liquid-based methods has some weaknesses like the complicated measurement protocols, long measurement time, and environmental problems with chemical disposal (Heister, 2014). Considering these limitations, the use of water to estimate SSA is a better alternative and has been previously applied (Newman, 1983; Puri & Murari, 1964). Estimation of SSA from water sorption or retention is often achieved by combing water vapor sorption isotherm (WSI) measurements with physically based (e.g., Tuller & Or, 2005) or empirical (e.g., Resurreccion et al., 2011) models. The isotherms represent the relationship between relative humidity (water activity) and the equilibrium soil–water content at a given temperature, obtained along an adsorption (wetting) or desorption (drying) path. Recent technological advances have led to faster, more detailed, and reliable measurements of the WSI. Arthur, Tuller, Moldrup, and de Jonge (2014) reported the great potential of an automated vapor sorption analyzer (VSA) for soil exploration, including estimations of clay content and SSA, as well as solute percolation threshold and cation exchange capacity. To estimate SSA, WSIs were often used in conjunction with different modeling approaches.

The Brunauer–Emmet–Teller (BET) model is a monolayer approach to estimate SSA, usually applied in conjunction with gas (N_2 or other gases) (Brunauer, Emmett, & Teller, 1938) and works well but only for nonswelling soils (Khorshidi, Lu, Akin, & Likos, 2017). The Guggenheim–Anderson–de Boer (GAB) model is similar to the BET equation but accounts for multilayer molecules relative to the bulk liquid. It presents a good alternative to the BET model and was reported to be accurate for both natural and swelling soils (Akin & Likos, 2017; Arthur, Tuller, Moldrup, & de Jonge, 2016). The Tuller–Or (TO) model is a physically based water film

Core Ideas

- A new method to estimate SSA by combining vis-NIRS and two WSI models is proposed.
- The vis-NIRS models are also compared with models where the SSA is determined with the EGME method.
- Three types of regression techniques including PLS, SVM, and ANN are tested.
- The combination of vis-NIRS with the WSI as a reference provides SSA estimates similar to the EGME method.

adsorption model parameterized with adsorption WSIs. The TO model is reported to be most suitable for soils exhibiting SSA values in the range of 5–200 $m^2\ g^{-1}$ (Akin & Likos, 2014; Arthur et al., 2013; Khorshidi et al., 2017; Leão & Tuller, 2014; Tuller & Or, 2005). However, it fails to accurately describe the drier parts of the adsorption isotherms (Resurreccion et al., 2011).

Visible near-infrared spectroscopy (vis-NIRS) is another promising alternative technique for SSA estimation. It is a versatile and robust analytical technique with a high repeatability and a demonstrated record of successful application to soil analysis. The vis-NIRS is based on the interaction of light with the soil sample under investigation. The output is a vis-NIR reflectance spectrum (400–2500 nm), represented as measured vis-NIR intensities vs. wavelength of electromagnetic radiation. The vis-NIR spectrum reflects the presence of chemical functional groups related to the mineral and organic composition of the sample, thus being relevant for the estimation of physical and chemical soil properties. It is a very efficient method (short measurement time and minimal sample preparation) that does not require chemicals and does not destroy the sample. With only one obtained spectrum, multiple soil properties can be determined (Pasquini, 2003). The vibrational modes in the vis-NIR region are, however, weak and typically cause broad and overlapping absorption features. In order to assign specific features to specific chemical components, multivariate calibrations are used (Martens & Næs, 1989). Different methods can be applied to correlate soil spectra with the soil constituents of interest. The most common include linear models such as principal component regression, partial least square (PLS) regression, multiple linear regression, and stepwise multiple linear regression (Soriano-Disla, Janik, Rossel, Macdonald, & McLaughlin, 2014). Nonlinear models include machine-learning techniques such as multivariate adaptive regression splines, artificial neural networks (ANN), regression trees, or

support vector machines (SVM) (Viscarra Rossel & Behrens, 2010). Although the application of PLS in soil spectroscopy is most prevalent in the literature, machine-learning algorithms have been reported to provide higher estimation accuracy for a range of soil properties (Viscarra Rossel & Behrens, 2010).

Extensive research efforts have been devoted in the last decade to using vis-NIRS in combination with multivariate techniques as a powerful means to overcome the time-consuming and often complicated classical analysis of both fundamental and functional soil properties (Hermansen et al., 2017; Katuwal et al., 2017; Knadel et al., 2016; Nocita et al., 2012; Paradelo et al., 2016; Pittaki et al., 2018, 2019; Viscarra Rossel et al., 2016). However, the application of vis-NIRS to SSA determination is still relatively rare. The few successful attempts to determine the SSA from vis-NIR spectra included the predictions of SSA obtained from the EGME method only (Ben-Dor & Banin, 1995; Ben-Dor, Heller, & Chudnovsky, 2008; Knadel et al., 2018).

To further investigate the applicability of vis-NIRS for SSA estimation, the objectives of this study are

(i) to test the feasibility of vis-NIRS for SSA estimation, where the SSA is determined with the TO (SSA_{TO}) and GAB (SSA_{GAB}) models parameterized with WSIs measured with a VSA, and using three types of regression techniques (PLS, ANN, and SVM),

(ii) to compare the generated vis-NIRS models for SSA with SSA models where SSA was estimated with the EGME method (SSA_{EGME}),

(iii) to investigate the effect of dataset subdivision according to EGME values on the performance of the vis-NIRS models.

2 | MATERIALS AND METHODS

2.1 | Investigated soil samples

A total of 270 soil samples (220 topsoils and 50 subsoils) were investigated in this study. The samples represent a wide range of soil types, mineralogies, and geographic origins that include Europe (Denmark, Germany, Spain, Norway, and Belgium; $N = 116$), North America (the United States; $N = 57$), South and Central Americas (Brazil, Colombia, Ecuador, Peru, Venezuela, Uruguay, Cuba, and Nicaragua; $N = 42$), Africa (Ghana, Nigeria, Mozambique, Zimbabwe, Cameroon, Ethiopia, South Africa, Kenya, and Côte d'Ivoire; $N = 38$), and Asia (India, Sri Lanka, Malaysia, and China; $N = 17$). In brief, the samples have been obtained from different locations in various countries. Some of the samples were extracted from agricultural fields with gradients in clay and/or organic C contents (Supplemental Table S1), others from different agroecological regions within a country, and some from the large soil

database of the International Soil Reference and Information Centre (Wageningen). Further descriptions of individual samples and their properties, soil type and sampling locations are provided in Supplemental Table S1.

2.2 | Reference soil measurements

All soil samples were air dried and sieved to 2 mm prior to the analyses described below. After removal of organic matter and carbonates, particle size fractions were determined with a combination of wet sieving and pipette or hydrometer methods (Gee & Or, 2002). The soil organic C (SOC) was either determined based on the principle of C oxidation at 1800 °C using an elemental analyzer with a thermal conductivity detector (Thermo Fisher Scientific) or by wet combustion using the Walkley–Black method (Nelson & Sommers, 1982). The SSA was determined in the laboratory via retention of EGME at monolayer coverage (Pennell, 2002) without organic C removal or ion saturation.

2.3 | Water vapor sorption measurements

Soil WSIs were obtained with a fully automated VSA (METER Group). The VSA system dries and wets the air-dry sample (∼3.5 g soil) and measures the water potential using a chilled-mirror dewpoint method. The sample mass is automatically recorded during the drying and wetting process with a high-precision magnetic force balance (Arthur et al., 2013; Likos, Lu, & Wenszel, 2011).

The isotherms were measured in dynamic dewpoint mode for adsorption and desorption for a water activity range from 0.03 to 0.93 and a temperature of 25 °C. The reference water content for all samples was calculated after oven drying at 105 °C for 48 h. For a detailed description of the VSA, interested readers are referred to Arthur et al. (2014).

2.3.1 | Tuller–Or model

The physically based TO model Equation 1 was parameterized with water adsorption data for the matric potential (ψ) range from −470 to −10 MPa (corresponding to the water activity range from 0.03 to 0.93). The TO model relates the equilibrium water content, M (kg kg^{-1}), to ψ (cm H$_2$O) and the SSA (m^2 kg^{-1}) as

$$M = \sqrt[3]{\frac{A_{svl}}{6\pi\rho_w g\psi}} SSA_{TO} \qquad (1)$$

where A_{svl} (J) is the Hamaker constant for solid–vapor interactions through the intervening liquid, ρ_w is the density of water (kg m^{-3}), and g is acceleration due to gravity (m s^{-2}). The

value of A_{svl} was set to -6×10^{-20} J, as suggested in Tuller and Or (2005) and Maček, Mauko, Mladenovič, Majes, and Petkovšek (2013).

2.3.2 | Guggenheim–Anderson–de Boer model

The GAB model relates the water activities to the equilibrium water contents (M, kg kg^{-1}) via three model parameters (M_0, C, and K):

$$M = \frac{M_{0G} C_G Ka_w}{[(1 - Ka_w)(1 - Ka_w + C_G Ka_w)]} \quad (2)$$

where M_0 (kg kg^{-1}) is the monolayer water content, C_G is an energy constant, and Ka_w represents the difference of free enthalpy of the water molecules in the pure liquid and the layers above the monolayer Equation 2. Since the GAB model can be parameterized with both adsorption and desorption data, here we applied desorption data. This was because the adsorption data are not always reproducible due to their sensitivity to initial water content, hydrophobicity, and stronger intermolecular forces than experienced for desorption (Johansen & Dunning, 1957; Lu & Khorshidi, 2015).

The SSA$_{GAB}$ was calculated with Equation 3 (Newman, 1983; Quirk & Murray, 1999):

$$SSA = \frac{M_0 N A}{w_M} \quad (3)$$

where M_0 is the monolayer water content (kg kg^{-1}) from the GAB equation, N is Avogadro's number (6.02×10^{23} mol^{-1}), A is the area covered by one water molecule (10.8×10^{-20} m^2), and w_M is the molecular weight of water (0.018 kg mol^{-1}).

2.4 | Vis-NIRS measurements

Spectral measurements were performed in the visible and near-infrared range (400–2500 nm) with a NIRS DS2500 spectrophotometer (FOSS) in a temperature- and humidity-controlled room (temperature of 23 °C, humidity of 48%). Air-dried and 2-mm-sieved soil samples (~50 g) were scanned in seven spots each through a quartz window of the sample holder. An average of the seven scans (absorbance spectrum (Abs) = [log(1/R)], where R is reflectance) was used further in the modeling phase.

2.5 | Datasets

Calibration models were generated to demonstrate the potential of vis-NIRS for SSA estimation for this diverse dataset and

were based on the full dataset, as well as on datasets obtained after subsetting, where the distribution of SSA values was considered. Due to skewness in the SSA$_{EGME}$ values (almost 70% of the samples exhibited SSA$_{EGME}$ values <100 m^2 g^{-1}), the data were divided into two subsets, with SSA$_{EGME}$ < 100 m^2 g^{-1} ($N = 180$) and SSA$_{EGME}$ > 100 m^2 g^{-1} ($N = 90$).

To ensure a representative selection of calibration sets for vis-NIRS modeling, a principal component analysis (PCA) was performed (Webster & Oliver, 2001) for spectral data of each dataset considered above, and the Kennard–Stone algorithm (Kennard & Stone, 1969) was applied to the scores of the first three principal components. The algorithm was set to select 80% of the samples for calibration, with the remaining 20% assigned to a validation dataset. This resulted in a calibration and a validation set for the entire dataset including 216 and 54 samples (validation samples were marked in gray in Supplemental Table S1), respectively, and four subsets considering the SSA distribution: calibration ($N = 144$) and validation (N = 36) subsets for the set with SSA$_{EGME}$ < 100 m^2 g^{-1}, and calibration ($N = 70$) and validation subsets ($N = 20$) for the set with SSA$_{EGME}$ > 100 m^2 g^{-1}. To avoid an issue with pseudoreplicates in the calibration and validation subsets (as in few cases that the samples with a gradient in SSA were obtained from one field), all field samples were kept in the calibration datasets.

2.6 | Multivariate data analysis

In order to derive information on soil constituents from the weak and broad absorptions in vis-NIR spectra, three types of regression techniques were used: PLS, ANN, and SVM. All of them were using calibration samples to generate models for SSA determined by the TO and GAB methods, and by the EGME method. Moreover, models for texture (clay, silt, and sand) and SOC were also generated (but only for the first calibration and validation approach on the entire dataset). The training of all calibration models was performed with a single 10-fold venetian blinds cross-validation. In this calibration method, 10% of the data were withheld and used to validate the calibration model built on the data of the remaining samples. This was repeated until all samples were left out once. All calibration models were further validated with the independent validation sets. Modeling was performed with the Matlab PLS Toolbox 8.7 (Eigenvector Research).

2.7 | Partial least squares regression

Partial least squares regression is one of the most commonly used regression methods that produced satisfactory calibration results for a variety of soil constituents. It models both the $\mathbf{X}$ (spectra) and $\mathbf{Y}$ (soil constituent of interest) matrices

simultaneously by compressing and regressing the data to find the latent variables (factors) in **X** that best predict the latent variables in **Y**. This regression technique reduces data dimensionality and noise and is computationally faster. It is used for highly collinear predictor variables. Here, PLS with a noniterative partial least square algorithm was applied (Martens & Næs, 1989; Wold, Sjöström, & Eriksson, 2001).

2.8 | Artificial neural networks

An ANN is a framework for a range of machine-learning algorithms designed to imitate the way a brain performs different tasks. It is a group of three layers of interconnected nodes (artificial neurons). The three layers include input (here, vis-NIR spectra), hidden (a layer between the input and output), and output (the property to be predicted). The nodes from one layer are connected with the nodes from the adjacent layer with a strength referred to as a weight. Each input within one layer is multiplied by a corresponding weight and is handled by an activation function, in the hidden layer, to produce an output. This is further used as an input in the next layer. The weights optimization is accomplished through a training procedure performed on a calibration set (Goldshleger, Chudnovsky, & Ben Dor, 2012). A feedforward ANN with a back-propagation neural network, which aims at minimizing the network error, was used. It finds the optimal number of iteration cycles by choosing the lowest RMSE of cross-validation based on the training data set and iteration values (here, 1–20) (Rumelhart, Hinton, & Williams, 1986). To shorten the computation time, the vis-NIR spectra were compressed using PLS regression and three principal components. Then ANN with two nodes in the first layer on the principal component scores was performed.

2.9 | Support vector machines

Support vector machines are nonlinear kernel-based learning methods. Here, the Gaussian radial basis function kernel type was used. Support vector machine regression trains nonlinear data by mapping them into a multidimensional kernel space and derives optimal bounds for regression (Vapnik, 1995). It defines the loss function, which ignores errors situated within a given distance of the true value. Models are built with a smaller set of representative observations close to the regression boundary (support vectors) (Suykens & Vandewalle, 1999). This algorithm requires model optimization by adjusting two parameters: ε (used values: 1.0, 0.1, 0.01), which is the upper tolerance on prediction errors, and C (11 values from 10^{-3} to 100, spaced uniformly on the log scale used), which determines the tradeoff between the model complexity and the degree to which deviations larger than ε are
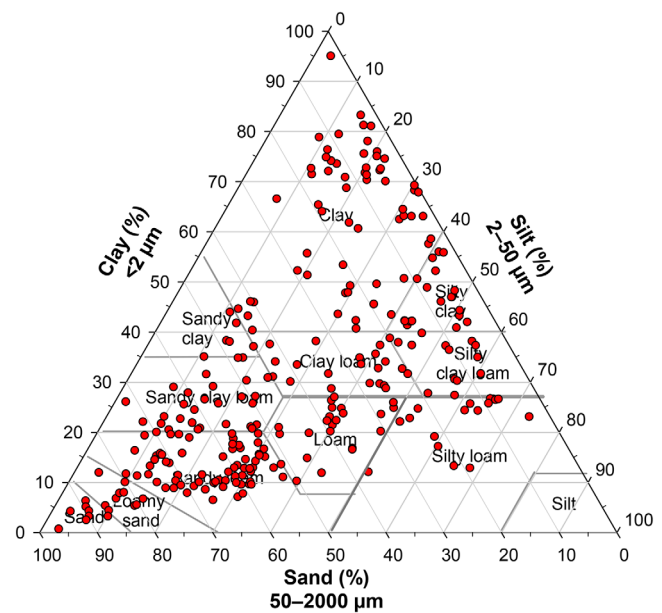

FIGURE 1 Distribution of investigated soil samples ($N = 270$) within the USDA soil textural triangle

tolerated. Additional details about the regression techniques can be found in Hastie, Tibshirani, and Friedman (2009).

The performance of the regression models was evaluated using the RMSE of cross-validation, the RMSE of prediction, and the R^2. Due to differences in the SSA range resulting from the use of different determination methods (TO or GAB model and the EGME method), the standardized RMSE was additionally calculated as SRMSE = RMSE/range, to enable the comparison between the performance of different models.

3 | RESULTS AND DISCUSSION

3.1 | Soils

The investigated samples represent a wide range of soil types (Figure 1), with clay contents ranging from 1 to 95% and sand contents ranging from 0 to 96%. The samples covered both mineral and more organic soils, with some containing >8% organic C (Table 1). Because of the diverse geographic origin of the considered soils, distinct differences in mineralogy are also expected. This high variability in soil properties of the investigated soil resulted in a wide range of SSA$_{EGME}$ values (6–445 m^2 g^{-1}) (Table 1).

3.2 | Water vapor sorption isotherms

Figure 2a presents three soils with different composition and thus varying SSA values. The WSIs (adsorption and desorption loops) follow a clear pattern, where the sample with large

T A B L E 1 General statistics of the investigated soil properties

	Property[a]						
	SSA_{TO}	SSA_{GAB}	SSA_{EGME}	Clay	Silt	Sand	SOC
Gen.stat [b]	$m^2\ g^{-1}$			%			
Mean	97 (97, 96)	107 (106, 112)	103 (101, 108)	32 (32, 31)	29 (27, 37)	39 (41, 33)	1.57 (1.66, 1.2)
Max.	374 (374, 370)	428 (417, 428)	444 (444, 400)	95 (95, 79)	68 (68, 62)	96 (96, 82)	8.42 (8.42, 3.5)
Min.	8 (8, 23)	7 (7, 24)	6 (6, 10)	1 (1, 5)	2 (2, 10)	0 (0, 3)	0 (0, 0.07)
SD	82 (81, 84)	99 (98, 107)	104 (103, 109)	21 (22, 18)	15 (15, 14)	26 (26, 24)	1.39 (1.49, 0.76)
Variance	6,673 (6,617, 7,024)	9,875 (9,537, 11,402)	10,906 (10,681, 11,989)	445 (475, 327)	239 (229, 208)	676 (690, 569)	1.92 (2.22, 0.58)
Skewness	2 (2,1)	2 (2, 1)	2 (2, 1)	1 (1, 1)	1 (1,0)	0 (0, 0)	1.80 (1.66, 0.74)
Q1	41 (42, 38)	37 (38, 35)	33 (32, 34)	15 (14, 17)	16 (15, 26)	14 (16, 11)	0.69 (0.7, 0.66)
Q3	126 (124, 130)	140 (135, 184)	132 (130, 160)	43 (44, 42)	40 (37, 48)	61 (62, 53)	1.97 (2.09, 1.51)

Note. The first value is for the entire dataset ($N = 270$), the first value in brackets is for the calibration dataset ($N = 216$), and the second value in the brackets is for the validation dataset ($N = 54$).

[a]SSA_{TO}, soil specific surface area (SSA) determined from vapor sorption isotherms using the Tuller–Or model; SSA_{GAB}, SSA determined from vapor sorption isotherms using a Guggenheim–Anderson–Boer model; SSA_{EGME}, SSA determined using ethylene glycol monoethyl ether method; SOC, soil organic C.

[b]Gen.stat, general statistics; Q1, the first quartile, Q3, the third quartile.
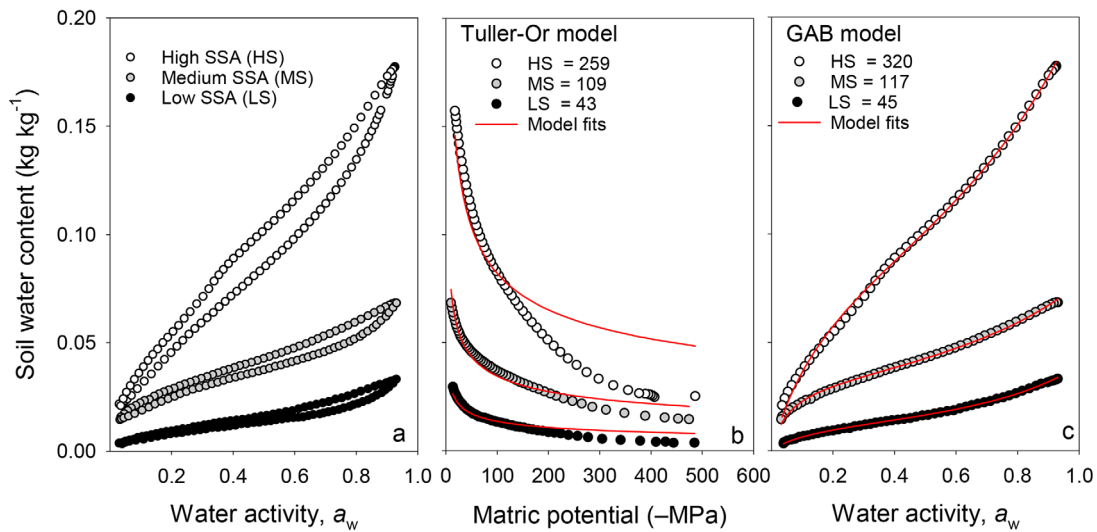

F I G U R E 2 (a) Example of measured water vapor sorption isotherms for three samples with different soil specific surface areas (SSAs), (b) fit of Tuller and Or (2005) model to the adsorption isotherms, and (c) fit of the Guggenheim–Andersen–de Boer (GAB) model to the desorption isotherms. The ethylene glycol monoethyl ether estimates of SSA for the high-, medium-, and low-SSA samples were 307, 111, and 45 $m^2\ g^{-1}$, respectively. The numbers in the legend of Panels b and c are the SSA estimates in $m^2\ g^{-1}$ from the two models

SSA had higher soil water sorption for any given water activity value. The fits of the TO model (fitted to the adsorption isotherms) and the GAB model (fitted to the desorption isotherms) reflect the same behavior for the three soils (Figures 2b and 2c). The GAB model predicted water content well, regardless of the soil type and water activity value (Figure 2c), whereas the TO model (Figure 2b) described the adsorption isotherms well only up to −200 MPa for the soil with the medium ($SSA_{EGME} = 111\ m^2\ g^{-1}$) and low

($SSA_{EGME} = 45\ m^2\ g^{-1}$) SSA values, and up to −120 MPa for the soil with the highest SSA value ($SSA_{EGME} = 307\ m^2\ g^{-1}$). Above these thresholds, a clear overprediction can be observed. This is in line with the previous findings, where the TO-predicted water contents were up to 50% higher than foreseen and were attributed to higher errors for the finer-textured soils (Arthur et al., 2013; Resurreccion et al., 2011). This is perhaps the reason why the SSA_{TO} estimated for large-surface-area samples was less than the SSA_{GAB} and
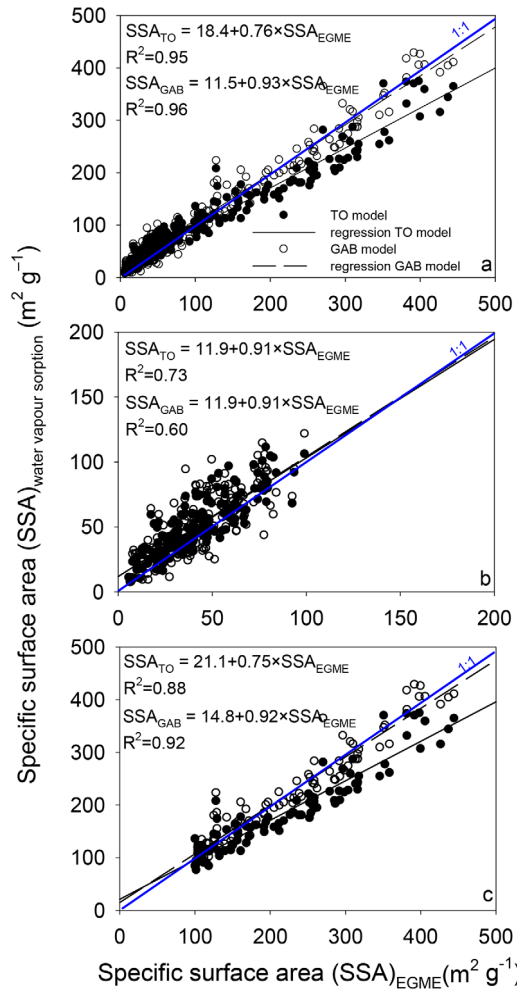
FIGURE 3 Relationships between soil specific surface areas (SSAs) determined with the ethylene glycol monoethyl ether (EGME) method, with SSA derived with the Tuller–Or (TO) and Guggenheim–Anderson–de Boer (GAB) models, for (a) the entire dataset ($N = 270$), (b) the subset with $SSA_{EGME} < 100$ m^2 g^{-1} ($N = 180$), and (c) the subset with $SSA_{EGME} > 100$ m^2 g^{-1} ($N = 90$). For all sets, $p < .001$ was reported for the regression analyses

SSA_{EGME}, with the maximum SSA values obtained being 374, 428, and 445 m^2 g^{-1}, respectively.

The correlations between the SSA_{TO} and SSA_{GAB} with the SSA_{EGME} values were very high ($R^2 = .95$ and .96, respectively) (Figure 3a). As discussed above, the TO model does not work optimally for soils with high SSA values and thus started deviating from the SSA_{EGME} values at ~150 m^2 g^{-1} (Figure 3a).

Due to skewed SSA_{EGME} values, the dataset was further divided into two subsets: $SSA_{EGME} < 100$ m^2 g^{-1} ($N = 180$) and $SSA_{EGME} > 100$ m^2 g^{-1} ($N = 90$) (Supplemental Tables S2 and S3). In general, lower correlations between the SSA_{TO} and SSA_{GAB} values with the SSA_{EGME} values were observed after subsetting, when compared with application of the full dataset (Figures 3b and 3c). The SSA values estimated with

both the TO and GAB models were larger for the subset with SSA_{EGME} values < 100 m^2 g^{-1} than the values obtained with the EGME method. In turn, lower values for the TO model for the set with values > 100 m^2 g^{-1} than the values obtained by the EGME method were obtained. Moreover, for the subset with the SSA_{EGME} values > 100 m^2 g^{-1}, higher correlations with SSA_{EGME} (R^2 of .88 and .92, for SSA_{TO} and SSA_{GAB}, respectively) than for the subset with the SSA_{EGME} values < 100 m^2 g^{-1} (R^2 of .73 and .60, for SSA_{TO} and SSA_{GAB}, respectively) were reported (Figures 3b and 3c).

3.3 | Vis-NIRS models

3.3.1 | Full dataset

For the full dataset, the best vis-NIRS calibration models for SSA_{TO} and SSA_{GAB} exhibited identical estimation accuracy to the SSA_{EGME} model (SRMSE = 0.10) (Figure 4). The best texture and SOC models obtained here had lower precision for calibration (average SRMSE of 0.18 for texture and 0.13 for SOC estimations) than those for the SSA models (Supplemental Table S4). Among the three regression techniques, SVM was the most accurate for estimating SSA_{TO}, clay, silt, sand, and SOC, and PLS generated the best results for SSA_{GAB} and SSA_{EGME}, whereas ANN showed the lowest estimation accuracy of the calibration models (Figure 4, Supplemental Table S4).

The independent validation of the developed calibration models for the three SSA estimates reflected the accuracy of the calibration models (Figure 5). The validation results from the best calibration model for SSA_{TO} slightly outperformed (SRMSE = 0.08) that of SSA_{GAB} (SRMSE = 0.10) but was similar to the SSA_{EGME} model (SRMSE = 0.09). High R^2 values (>.89) were obtained for all SSA estimations. The SSA_{EGME} validation results exhibit higher accuracy than obtained in Knadel et al., 2018 (SRMSE = 0.13), who validated with a set representing a smaller range of SSA_{EGME} values (4–116 m^2 g^{-1}), but also lower SD values (SD = 27) than the validation set used in this study (range: 10–392 m^2 g^{-1}, SD = 94) (Table 1).

When comparing the performance of these models with the best texture and SOC validation results (Supplemental Table S4), the SSA models showed better estimation accuracy than the texture and SOC models, which had average SRMSEs of prediction of 0.49 and 0.66, respectively.

3.3.2 | Subsets according to SSA_{EGME} values

To test how the range of the SSA affects the performance of SSA models, the same modeling analysis on the sets with $SSA_{EGME} < 100$ m^2 g^{-1} ($N = 180$) and $SSA_{EGME} > 100$ m^2
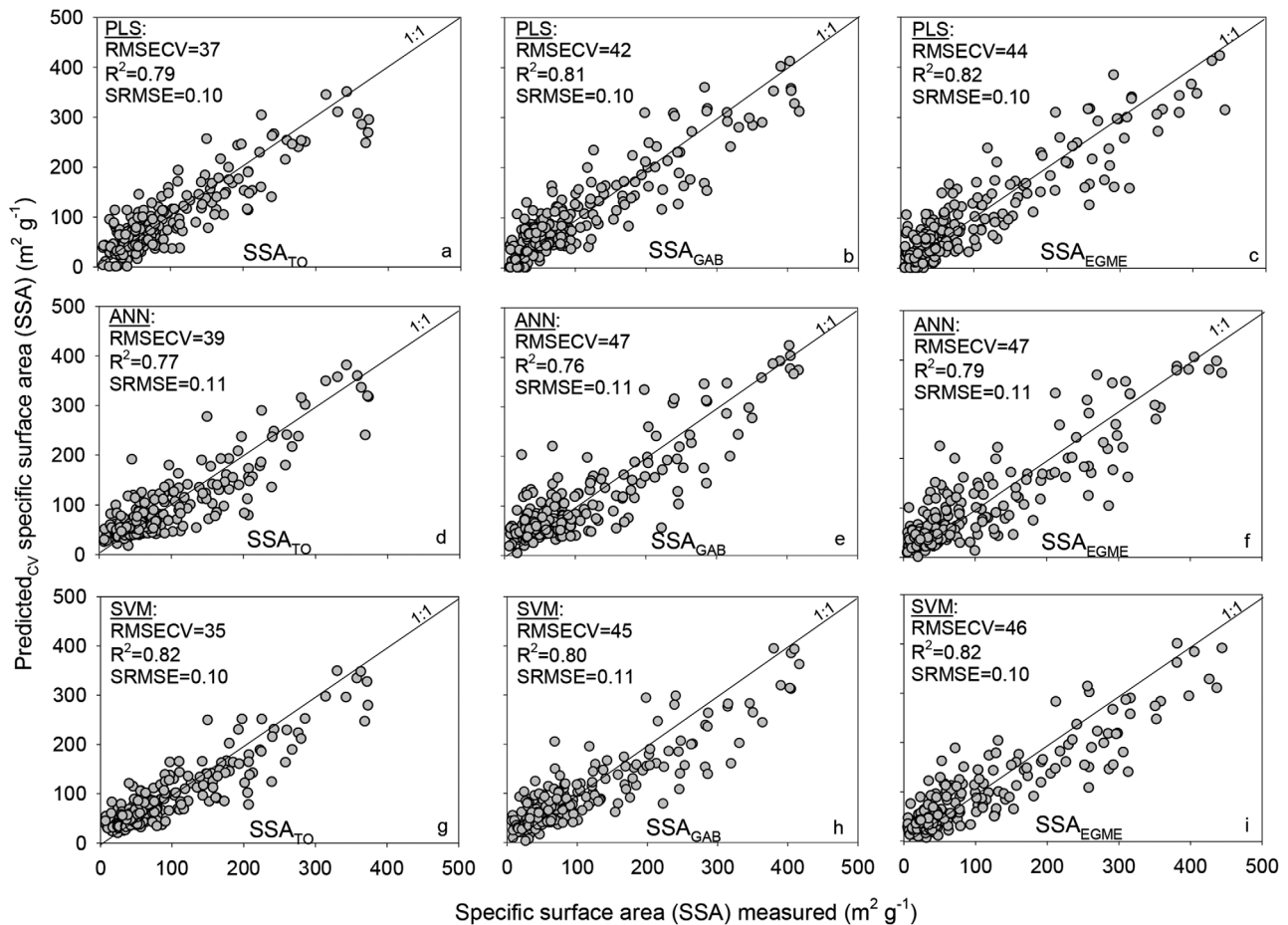
FIGURE 4 Visible–near-infrared spectroscopy calibration results ($N = 216$) for the soil specific surface area (SSA) presented as predicted (cross-validation [cv]) vs. measured for the Tuller–Or (TO) and Guggenheim–Anderson–de Boer (GAB) models, and the ethylene glycol monoethyl ether (EGME) method generated using partial least squares (PLS), artificial neural networks (ANN), and support vector machine (SVM) regression techniques. SRMSE = RMSE/range

g^{-1} ($N = 90$) was performed (Supplemental Tables S2 and S3). Detailed results of calibration and validation for both subsets are presented in Supplemental Figures S1–S4. Both calibration and validation results for the best SSA models of the subset with SSA_{EGME} values < 100 m^2 g^{-1} exhibited much lower estimation accuracy with higher SRMSE (0.18 on average) and lower R^2 values (.22–.45) than for the set with SSA_{EGME} values > 100 m^2 g^{-1} (on average, SRMSE of 0.14 and R^2 values between .63 and .78). The discrepancies in model performance for the two subsets can be related to the effects of variation in SSA values themselves, the organo-mineral composition, and their interactions. The subset with SSA_{EGME} values > 100 m^2 g^{-1} presents higher standard deviations (Supplemental Table S3), and this was previously related to elevated R^2 values (Stenberg, 2010). Additionally, this subset includes samples with the highest clay contents (on average, 51%), whereas the subset with the SSA_{EGME} values < 100 m^2 g^{-1} includes soils with an average clay content of 21%. Higher clay content results in more pronounced absorptions from molecular bonds related

to clay minerals, but also to SOC (Stevens, Nocita, Tóth, Montanarella, & Van Wesemael, 2013), the two soil properties greatly affecting SSA (Knadel et al., 2018). Thus, we found improved SSA model performance for the set with SSA_{EGME} values > 100 m^2 g^{-1}, which was also characterized by a higher clay content. In contrary, the subset with SSA_{EGME} values < 100 m^2 g^{-1} represented mostly sandy soils (average sand content of 49%). Therefore, weak signals from clay minerals in the vis-NIR range were present. Moreover, high sand content increases light scattering and was reported to have a negative effect on model performance of SOC (Stenberg, 2010; Stevens et al., 2013).

Knadel et al. (2018) showed that, aside from the differences in texture and SOC content, the complexation status of SOC also affects the vis-NIRS estimation of SSA. The subset with SSA_{EGME} values > 100 m^2 g^{-1} represents soils with the capacity of clay to complex SOC (clay/SOC ratio, defined by Dexter et al., 2008, as $n = 10$) with n values > 10 (Supplemental Figure S5), meaning that soils unsaturated with SOC are present and all SOC is in complexed form. The subset
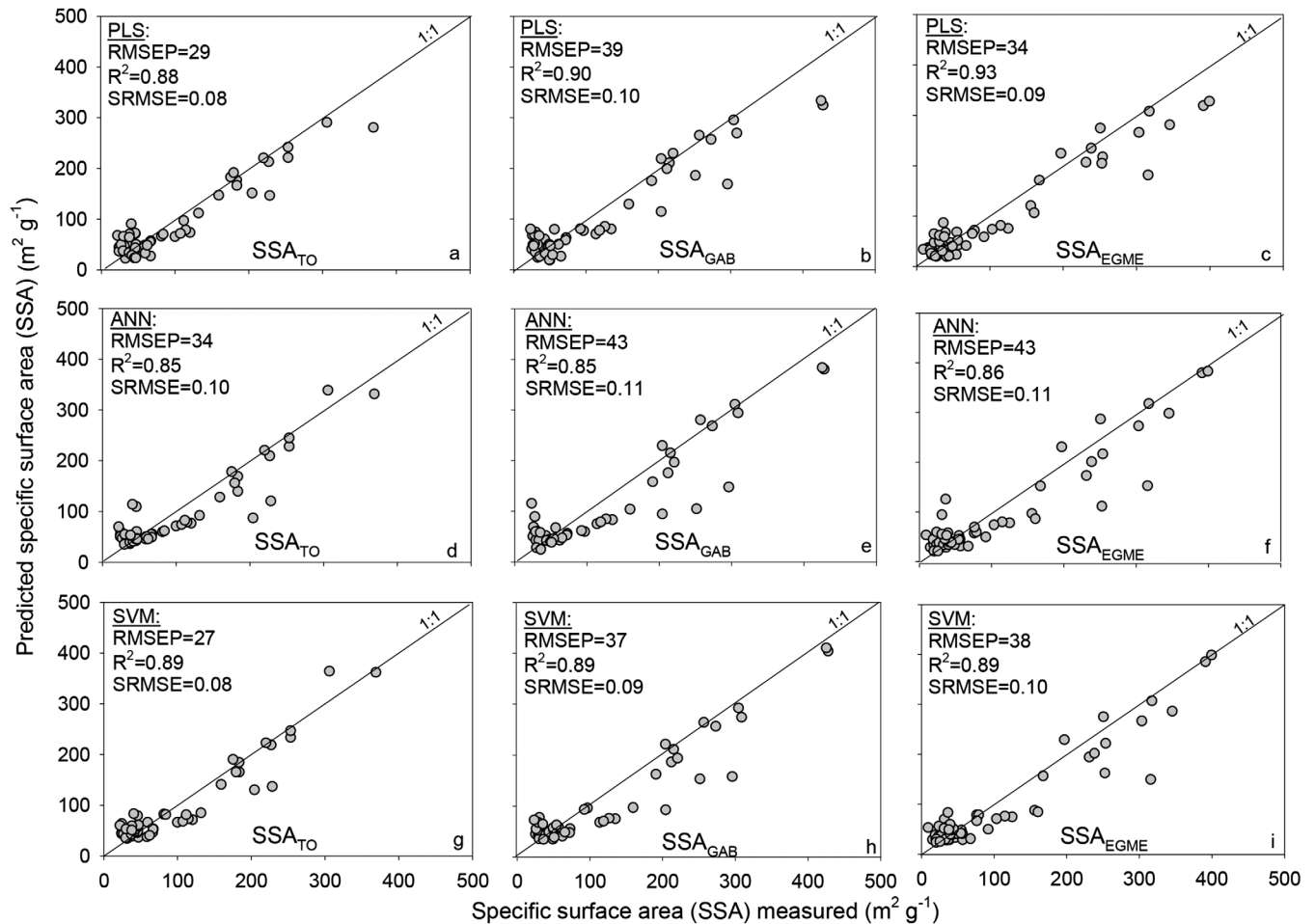
FIGURE 5 Visible–near-infrared spectroscopy validation results ($N = 54$) for the soil specific surface area (SSA) presented as predicted versus measured for the Tuller–Or (TO) and Guggenheim–Anderson–de Boer (GAB) models and the ethylene glycol monoethyl ether (EGME) method generated using partial least squares (PLS), artificial neural networks (ANN), and support vector machine (SVM) regression techniques. SRMSE = RMSE/range. RMSEP, RMSE of prediction

with the SSA_{EGME} values < 100 m^2 g^{-1}, in turn, represents soils with both noncomplexed and complexed forms of SOC ($10 < n > 10$) (Supplemental Figure S5). Therefore, the mineral surfaces of the samples with noncomplexed SOC have the potential to be coated with SOC (Knadel et al., 2018), which can potentially mask a portion of the SSA. This, together with the fact that both complexation forms were present, as well as the above-listed confounding effects of other soil constituents (like different clay mineralogy and negative effect of sand fractions) and the range of SSA values, potentially led to degraded SSA models for the subset with SSA_{EGME} values < 100 m^2 g^{-1}.

The SRMSE values obtained from the calibration models for each regression technique (PLS, ANN, SVM) for the full dataset and SSA subsets, and for each measure of SSA, are presented in Figure 6. For the models based on the full dataset, PLS resulted in the lowest errors for SSA_{GAB} and SSA_{EGME} estimation, whereas SVM provided best estimates for SSA_{TO}. After subsetting the data according to SSA_{EGME}

values, SVM performed better than the two remaining techniques for all three SSA estimates. Thus, on average, SVM resulted in higher accuracy. This points to an advantage of using machine-learning techniques over PLS and is in line with other studies where the application of machine-learning algorithms such as SVM outperformed PLS regression for soil property determination (Goldshleger et al., 2012; Kuang, Tekin, & Mouazen, 2015; Morellos et al., 2016; Tekin, Zeynal, & Mouazen, 2011; Viscarra Rossel & Behrens, 2010). However, the differences in the values of SRMSE among the different techniques were small and dependent on the dataset.

The comparison of results based on different validation datasets is somewhat problematic. Each of these sets consisted of different samples, a different total number of samples, and samples covering different ranges of SSA values. Therefore, even when using a standardized error, the comparison is not optimal. Thus, in order to perform a fair comparison, common validation samples existing in the validation set for a full dataset ($N = 54$), as well as in one of the validation sets for the
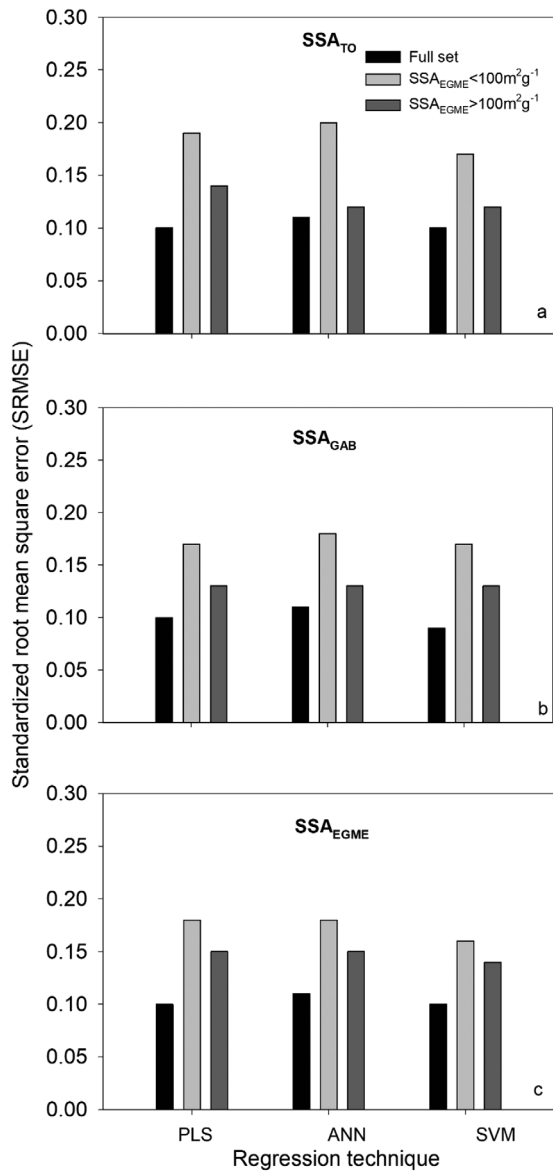
FIGURE 6 Comparison of the SRMSE (RMSE/ range) for the (a) Tuller–Or (TO), (b) Guggenheim–Anderson–de Boer (GAB), and (c) ethylene glycol monoethyl ether (EGME) estimates of soil specific surface area (SSA) based on visible–near-infrared spectroscopy modeling results using calibration datasets for all samples ($N = 54$), samples with $SSA_{EGME} < 100$ m^2 g^{-1} ($N = 36$), and samples with $SSA_{EGME} > 100$ m^2 g^{-1} ($N = 20$), generated with partial least squares (PLS), artificial neural networks (ANN), and support vector machine (SVM) regression techniques
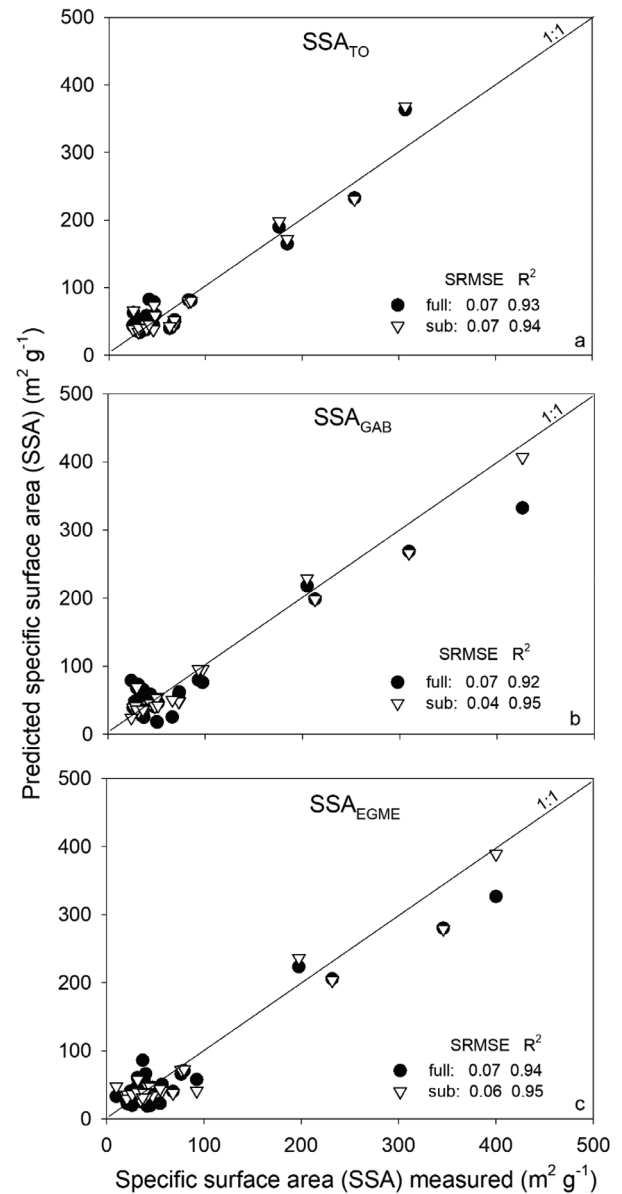
FIGURE 7 Comparison of model performance (standardized room mean square error [SRMSE] = RMSEP/range and R^2) for the (a) Tuller–Or (TO), (b) Guggenheim–Anderson–de Boer (GAB), and (c) ethylene glycol monoethyl ether (EGME) estimates of soil specific surface area (SSA) based on visible–near-infrared spectroscopy calibration models for common validation samples ($N = 25$) occurring in the full dataset (full, black circle) and subsets according to EGME values (sub, open triangle)

subsets with the $SSA_{EGME} < 100$ m^2 g^{-1} ($N = 36$) and the subset with the $SSA_{EGME} > 100$ m^2 g^{-1} ($N = 20$), were extracted. In total 28 common samples were found, and their estimations from the three calibration approaches were compared (Figure 7). In general, higher estimation accuracy was obtained after subsetting the data, with the greatest improvement seen for SSA_{GAB} (SRMSE of 0.07 and R^2 of .92 before subsetting, and SRMSE of 0.04 and R^2 of .95 after

subsetting). Nevertheless, there were no significant differences between the subsetting methods when the differences between the reference values and the predicted SSA values were compared for each SSA estimate (Mann–Whitney rank sum test, $P = .617$ for $SSA_{TO,}$ $P = .7$ for SSA_{GAB}, and $P = .8$ for SSA_{EGME}). Moreover, the estimation accuracy of vis-NIRS models for the SSA obtained by the two WSIs models (TO and GAB with an average SRMSE of 0.07 and 0.06,

respectively) and for the 28 common samples was nearly identical to that of vis-NIRS models for SSA$_{EGME}$ (average SRMSE of 0.06).

4 | CONCLUSIONS

In this study, vis-NIRS combined with different modeling techniques (PLS, ANN, and SVM) was applied to estimate SSA determined with two WSI-based models (SSA$_{TO}$ and SSA$_{GAB}$) for a heterogeneous soil sample set. The vis-NIRS SSA estimates were successful and indicated a similar estimation ability to a vis-NIRS model of SSA determined with the often-used EGME method. Furthermore, the performance of the models was mainly dependent on the range and variation in SSA values, as well as the organo-mineral composition and its interactions. However, no significant differences among the performance of calibration models, based on the entire dataset and the subsets in regards to SSA$_{EGME}$ values, were found for common validation samples. Moreover, in most cases, the application of SVM technique in the vis-NIRS modeling resulted in the best performance, yet the differences among the three types of regression techniques tested were small.

The elevated interest in SSA, which governs numerous soil processes and behaviors, calls for rapid, more accurate, and repeatable alternative methods for its determination. Given the results from this study, we suggest a combination of vis-NIRS, known for its reliable results, and the WSI as a reference technique for training vis-NIRS models, which does not involve the use of chemicals and provides SSA estimations similar to the EGME method. Although no significant differences in the estimation of SSA from vis-NIRS based on TO and GAB models have been observed, we recommend the use of the latter, as it is known to predict water contents well regardless of the soil type and water activity value. The performance of regression techniques applied, as well as spectral preprocessing methods, is usually dataset dependent, and we suggest testing different methods including both linear and nonlinear techniques to find the best option for the dataset under consideration.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ORCID
Maria Knadel https://orcid.org/0000-0001-7539-6191
Hafeez Ur Rehman https://orcid.org/0000-0002-7400-4026
Emmanuel Arthur https://orcid.org/0000-0002-0788-0712

REFERENCES
Akin, I. D., & Likos, W. J. (2014). Specific surface area of clay using water vapor and EGME sorption methods. *Geotechnical Testing Journal*, 37, 1016–1027. https://doi.org/10.1520/GTJ20140064

Akin, I. D., & Likos, W. J. (2017). Evaluation of isotherm models for water vapor sorption behavior of expansive clays. *Journal of Performance of Constructed Facilities*, 31(1). https://doi.org/10.1061/(ASCE)CF.1943-5509.0000899

Amali, S., Petersen, L. W., & Rolston, D. E. (1994). Modeling multicomponent volatile organic and water-vapor adsorption on soils. *Journal of Hazardous Materials*, 36, 89–108. https://doi.org/10.1016/0304-3894(93)E0082-D

Arthur, E., Tuller, M., Moldrup, P., & de Jonge, L. W. (2014). Rapid and fully automated measurement of water vapor sorption isotherms: New opportunities for vadose zone research. *Vadose Zone Journal*, 13(1). https://doi.org/10.2136/vzj2013.10.0185

Arthur, E., Tuller, M., Moldrup, P., & de Jonge, L. W. (2016). Evaluation of theoretical and empirical water vapor sorption isotherm models for soils. *Water Resources Research*, 52, 190–205. https://doi.org/10.1002/2015WR017681

Arthur, E., Tuller, M., Moldrup, P., Greve, M. H., Knadel, M., & de Jonge, L. W. (2018). Applicability of the GAB water vapour sorption model for estimation of soil specific surface area. *European Journal of Soil Science*, 69, 245–255. https://doi.org/10.1111/ejss.12524

Arthur, E., Tuller, M., Moldrup, P., Resurreccion, A. C., Meding, M. S., & Kawamoto, K. (2013). Soil specific surface area and non-singularity of soil-water retention at low saturations. *Soil Science Society of America Journal*, 77, 43–53. https://doi.org/10.2136/sssaj2012.0262

Ben-Dor, E., & Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal*, 59, 364–372. https://doi.org/10.2136/sssaj1995.03615995005900020014x

Ben-Dor, E., Heller, D., & Chudnovsky, A. (2008). A novel method of classifying soil profiles in the field using optical means. *Soil Science Society of America Journal*, 72, 1113–1123. https://doi.org/10.2136/sssaj2006.0059

Borkovec, M., Wu, Q., Degovics, G., Laggner, P., & Sticher, H. (1993). Surface-area and size distributions of soil particles. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 73, 65–76. https://doi.org/10.1016/0927-7757(93)80007-2

Brunauer, S., Emmett, P. H., & Teller, E. (1938). Adsorption of gases in multimolecular layers. *Journal of the American Chemical Society*, 60, 309–319. https://doi.org/10.1021/ja01269a023

Cerato, A., & Lutenegger, A. (2002). Determination of surface area of fine-grained soils by the ethylene glycol monoethyl ether (EGME) method. *Geotechnical Testing Journal*, 25, 315–321. https://doi.org/10.1520/GTJ11087J

de Jonge, H., de Jonge, L. W., & Mittelmeijer-Hazeleger, M. C. (2000). The microporous structure of organic and mineral soil materials. *Soil Science*, 165, 99–108.

de Jonge, H., & Mittelmeijer-Hazeleger, M. C. (1996). Adsorption of CO_2 and N_2 on soil organic matter: nature of porosity, surface area and diffusion mechanisms. *Environmental Science & Technology*, 30, 408–413. https://doi.org/10.1021/es950043t

Dexter, A. R., Richard, G., Arrouays, D., Czyz, E. A., Jolivet, C. & Duval, O. (2008). Complexed organic matter controls soil physical properties. *Geoderma*, *144*, 620–627. https://doi.org/10.1016/j.geoderma.2008.01.022

Gee, G. W., & Or, D. (2002). Particle-size analysis. In J. H. Dane & G. C. Topp (Eds.), *Methods of soil analysis: Part 4. Physical methods* (pp. 255–293). Madison, WI: SSSA. https://doi.org/10.2136/sssabookser5.4.c12

Goldshleger, N., Chudnovsky, A., & Ben Dor, E. (2012). Using reflectance spectroscopy and artificial neural network to assess water infiltration rate into the soil profile. *Applied and Environmental Soil Science*, 2012. https://doi.org/10.1155/2012/439567

Hang, P. T., & Brindley, G. W. (1970). Methylene blue absorption by clay minerals. Determination of surface areas and cation exchange capacities (clay organic studies XVIII). *Clays Clay Miner*, *18*, 203–212. https://doi.org/10.1346/CCMN.1970.0180404

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction.* New York: Springer. https://doi.org/10.1007/978-0-387-84858-7

Heister, K. (2014). The measurement of the specific surface area of soils by gas and polar liquid adsorption methods: Limitations and potentials. *Geoderma*, *216*, 75–87. https://doi.org/10.1016/j.geoderma.2013.10.015

Hermansen, C., Knadel, M., Moldrup, P., Greve, M., Karup Jensen, D., & de Jonge, L. W. (2017). Complete soil texture is accurately predicted by visible near-infrared spectroscopy. *Soil Science Society of America Journal*, *81*, 758–769. https://doi.org/10.2136/sssaj2017.02.0066

Johansen, R. T., & Dunning, H. N. (1957). Water-vapor adsorption on clays. *Clays Clay Miner*, *6*, 249–258. https://doi.org/10.1346/CCMN.1957.0060119

Katuwal, S., Hermansen, C., Knadel, M., Moldrup, P., Greve, M. G., & de Jonge, L. W. (2017). Combining X-ray computed tomography and visible near-infrared spectroscopy for prediction of soil structural properties. *Vadose Zone Journal*, *17*(1). https://doi.org/10.2136/vzj2016.06.0054

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*, 137–148. https://doi.org/10.2307/1266770

Khorshidi, M., Lu, N., Akin, I. D., & Likos, W. J. (2017). Intrinsic relationship between specific surface area and soil water retention. *Journal of Geotechnical and Geoenvironmental Engineering*, *143*(1). https://doi.org/10.1061/(ASCE)GT.1943-5606.0001572

Kim, K. C., Yoon, T., & Bae, Y. (2016). Applicability of using CO_2 adsorption isotherms to determine BET surface areas of microporous materials. *Microporous Mesoporous Materials*, *224*, 294–301. https://doi.org/10.1016/j.micromeso.2016.01.003

Knadel, M., Arthur, E., Jensen, P. W., Moldrup, P., Greve, M. H., Pittaki, Z., & de Jonge, L. W. (2018). Soil specific surface area determination by visible near-infrared spectroscopy. *Soil Science Society of America Journal*, *82*, 1046–1056. https://doi.org/10.2136/sssaj2018.03.0093

Knadel, M., Masis-Malendez, F., de Jonge, L. W., Moldrup, P., Arthur, E., & Greve, M. H. (2016). Assessing soil water repellency of a sandy field with visible near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *24*, 215–224. https://doi.org/10.1255/jnirs.1188

Kuang, B., Tekin, Y., & Mouazen, A. M. (2015). Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil & Tillage Research*, *146*, 243–252. https://doi.org/10.1016/j.still.2014.11.002

Leão, T. P., & Tuller, M. (2014). Relating soil specific surface area, water film thickness, and water vapor adsorption. *Water Resources Research*, *50*, 7873–7885. https://doi.org/10.1002/2013WR014941

Likos, W. J., Lu, N., & Wenszel, W. (2011). Performance of a dynamic dew point method for moisture isotherms of clays. *Geotechnical Testing Journal*, *34*, 373–382. https://doi.org/10.1520/GTJ102901

Lu, N., & Khorshidi, M. (2015). Mechanisms for soil-water retention and hysteresis at high suction range. *Journal of Geotechnical and Geoenvironmental Engineering*, *141*(8). https://doi.org/10.1061/(ASCE)GT.1943-5606.0001325

Maček, M., Mauko, A., Mladenovič, A., Majes, B., & Petkovšek, A. (2013). A comparison of methods used to characterize the soil specific surface area of clays. *Applied Clay Science*, *83–84*, 144–152. https://doi.org/10.1016/j.clay.2013.08.026

Martens, H., & Næs, T. (1989). *Multivariate calibration* New York: John Wiley & Sons.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., … Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, 104–116. https://doi.org/10.1016/j.biosystemseng.2016.04.018

Nelson, D. W., & Sommers, L. E. (1982). Total carbon, organic carbon, and organic matter. In A. L. Page (Ed.), *Methods of soil analysis: Part 2. Chemical and microbiological properties* Madison, WI: ASA and SSSA. https://doi.org/10.2134/agronmonogr9.2.2ed.c29

Newman, A. C. D. (1983). The specific surface of soils determined by water sorption. *European Journal of Soil Science*, *34*, 23–32. https://doi.org/10.1111/j.1365-2389.1983.tb00809.x

Nocita, M., Stevens, A., Van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthés, B., … Wetterlind, J. (2012). Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in Agronomy*, *132*, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002

Paradelo, M., Hermansen, C., Knadel, M., Moldrup, P., Greve, M. H., & de Jonge, L. W. (2016). Field-scale predictions of soil contaminant sorption using visible–near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *24*, 281–291. https://doi.org/10.1255/jnirs.1228

Pasquini, C. (2003). Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, *14*, 198–219. http://doi.org/10.1590/S0103-50532003000200006

Pennell, K. D. (2002). Specific surface area. In J. H. Dane & G. C. Topp (Eds.) *Methods of soil analysis, Part 4: Physical methods* (pp. 295–315). Madison, WI: SSSA. https://doi.org/10.2136/sssabookser5.4.c13

Petersen, L. W., Moldrup, P., Jacobsen, O. H., & Rolston, D. E. (1996). Relations between specific surface area and soil physical and chemical properties. *Soil Science*, *161*, 9–21. https://doi.org/10.1097/00010694-199601000-00003

Pittaki, Z., Arthur, E., Moldrup, P., Knadel, M., Iversen, B. V., Norgaard, T., & de Jonge, L. W. (2019). *Predicting the Campbell-Shiozawa soil-water retention function for the dry region: Comparing visible-near infrared spectroscopy with classical pedotransfer function.* Poster 1628. Poster presented at the annual meeting of the SSSA, San Diego, CA.

Pittaki, 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*. https://doi.org/10.2136/vzj2017.09.0169

Puri, B. R., & Murari, K. (1964). Studies in surface area measurements of soils: 2. Surface area from a single point on the water isotherm. *Soil Science*, *97*, 341–343.

Quirk, J. P., & Murray, R. S. (1999). Appraisal of the ethylene glycol monoethyl ether method for measuring hydratable surface area of clays and soils. *Soil Science Society of America Journal*, *63*, 839–849. https://doi.org/10.2136/sssaj1999.634839x

Resurreccion, A. C., Moldrup, P., Tuller, M., Ferre, T. P. A., Kawamoto, K., & Komatsu, T. (2011). Relationship between specific surface area and the dry end of the water retention curve for soils with varying clay andorganic carbon contents. *Water Resources Research*, *47*. https://doi.org/10.1029/2010WR010229

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In E. R. David, L. M. James, & C. P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Soriano-Disla, J., Janik, L. J., Rossel, R. A. 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*, *49*, 139–186. https://doi.org/10.1080/05704928.2013.811081

Stenberg, B. (2010). Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon. *Geoderma*, *158*, 15–22. https://doi.org/10.1016/j.geoderma.2010.04.008

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & Van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PLoS ONE*, *8*(6). https://doi.org/10.1371/journal.pone.0066409

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*, 293–300. https://doi.org/10.1023/A:1018628609742

Tekin, Y., Zeynal, T., & Mouazen, A. M. (2011). Effect of moisture content on prediction of organic carbon and pH using visible and near-infrared spectroscopy. *Soil Science Society of America Journal*, *76*, 188–198. https://doi.org/10.2136/sssaj2011.0021

Tuller, M., & Or, D. (2005). Water films and scaling of soil characteristic curves at low water contents. *Water Resources Research*, *41*. https://doi.org/10.1029/2005WR004142

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*, 45–54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra Rossel, R. A., Behrens, T., Ben Dor, E., Brown, D. J., Dematte, J. A. M., Shepherd, K.D., … Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, *155*, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Webster, R., & Oliver, M. A. (2001). *Geostatistics for environmental scientists*. Chichester, UK: John Wiley & Sons.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.