Nearest Neighbor Versus Regression Approach: Effect of Performance Measures, Calibration Set Size, and Sampling Method on Soil Organic Carbon Prediction Using VNIR Lab Spectroscopy

Chirag Rajendra Ternikar, Cécile Gomez, Debsunder Dutta, and D. Nagesh Kumar, *Senior Member, IEEE*

***Abstract*—Soil organic carbon (SOC) plays a critical role in soil health, agricultural productivity, and ecosystem functioning, making accurate SOC estimations essential for sustainable land management and climate change mitigation. Visible and near-infrared spectroscopy has emerged as a promising, nondestructive, and cost-effective method for SOC estimation. This study evaluates the performance of nine nearest neighbor (NN) models and the partial least squares regression (*PLSR*) model to estimate SOC using the global open soil spectral library data. Detailed error analyses and the use of mean absolute error (MAE) as performance metric revealed differences in model performance that traditional metrics like R^2, RMSE, and ratio of performance to deviation alone fail to capture. Error correlation analysis further indicated that *o_plsd* (optimized partial least squares distance, one of the NN models) and *PLSR* provide structurally independent insights, while certain pairs of NN models (*pcad – plsd* and *o_plsd – o_pcad*) yield redundant information. Among the ten models tested, *o_plsd* model outperformed *PLSR* by leveraging local data density, exhibiting lower MAE (1.79% versus 2.36%) but was more sensitive to reduction in calibration set size. In contrast, *PLSR* demonstrated better generalizability with less sensitivity to calibration size variation, but relatively higher sensitivity to the choice of sampling method. Future research should focus on strategies to improve computational efficiency of NN models. The findings highlight the importance of performance metric selection and calibration strategy in large-scale SOC modeling. These results have practical implications for improving SOC prediction models and designing efficient hybrid approaches for large, heterogeneous soil datasets.**

***Index Terms*—Error correlation analysis, nearest neighbor (NN) models, open soil spectral library, partial least squares regression (PLSR), soil organic carbon (SOC), structural independence, visible and near-infrared (VNIR) lab spectroscopy.**

Received 9 May 2025; revised 7 August 2025 and 20 August 2025; accepted 17 September 2025. Date of publication 29 September 2025; date of current version 17 October 2025. The work of Chirag Rajendra Ternikar was supported by the MHRD Fellowship provided by Ministry of Education, Government of India and Fellowship received from the Central Water Commission, Ministry of Jal Shakti, Government of India (GOI) under Dam Rehabilitation & Improvement program. *(Corresponding author: D. Nagesh Kumar.)*

Chirag Rajendra Ternikar and Debsunder Dutta are with the Department of Civil Engineering, Indian Institute of Science, Bengaluru 560012, India.

Cécile Gomez is with the LISAH, University of Montpellier, IRD, INRAE, Institut Agro, AgroParisTech, 34090 Montpellier, France, and also with the Indo-French Cell for Water Sciences, IRD, Indian Institute of Science, Bengaluru 560012, India.

D. Nagesh Kumar was with the Lyles School of Civil and Construction Engineering, Purdue University, West Lafayette, IN 47907 USA. He is now with the Department of Civil Engineering & Divecha Centre for Climate Change, Indian Institute of Science, Bengaluru 560012, India (e-mail: nagesh@iisc.ac.in).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2025.3615516, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2025.3615516

I. Introduction

SOIL, as the fundamental substrate supporting terrestrial life, is more than a physical support structure; it stores nutrients, regulates water availability, influences productivity, and hosts diverse microbial communities [1], [2]. Central to many of these functions is soil organic carbon (SOC), derived from decomposing plant and animal residues, which influences soil structure, water retention, and nutrient dynamics. It plays a significant role in soil health, fertility, and overall ecosystem functioning [3], [4], [5]. Monitoring SOC contents is essential for sustainable land management, assessing soil fertility, and mitigating climate change, as SOC is a major reservoir of carbon in terrestrial ecosystems [6].

Numerous methods for measuring SOC content are documented in the literature (see Appendix Section A). Traditional laboratory techniques, such as the Walkley–Black method [7] and loss on ignition method [8], provide precise estimates but are labor intensive, use chemical agents and prove impractical for large-scale assessments [9]. To overcome these limitations, visible and near-infrared (VNIR) spectroscopy has emerged as a promising, rapid, nondestructive and cost-efficient method [10], [11], [12]. Although VNIR based SOC predictions show varying degrees of success [13], [14], [15], implying that calibrations are not always guaranteed, thus warranting further investigation. A key aspect for successful SOC prediction is the representativeness of the calibration dataset. These datasets are mostly site-dependent, prompting the development of numerous local scale VNIR soil spectral libraries (SSL) adapted to specific regions [16], [17], [18], [19], [20], [21], [22]. In parallel, global scale VNIR SSLs have also been compiled to enhance broader applicability [23], [24], [25], [26], [27], [28]. Analyzing these large and global libraries presents significant challenges including spectral complexity, nonlinearity, large datasets, and the curse of dimensionality, requiring advanced modelling approaches.

Over time, the mathematical models to analyze such datasets have evolved from linear models, such as multiple linear regression (MLR) and partial least squares regression (PLSR) to more complex, nonlinear machine learning techniques (details

in Table III of Appendix A). PLSR remains a widely used linear model in soil spectroscopy [13], [24], but it struggles with datasets characterized by substantial nonlinearity and complexity. In contrast, nearest neighbor (NN) models, a subset of machine learning models, have gained popularity for their ability to leverage local spectral similarities and handle complex data structures [29], [30]. NN models involve finding the most similar neighbor based on distance metrics. The proximity between two objects is inversely related to their distance, with shorter distances indicating greater similarity between samples [29]. By leveraging the spectral signatures obtained from VNIR spectroscopy, the NN approach provides an efficient, data-driven approach for predicting SOC content [30]. While NN models show promise in capturing the nonlinear patterns present in large and heterogeneous SSLs, they have not been thoroughly evaluated using a global, diverse dataset like OSSL. This gap in the literature warrants a closer examination of NN models under real-world, large-scale conditions.

Many spectroscopy studies rely on metrics like coefficient of determination (R^2), root mean squared error (RMSE), ratio of performance to deviation (RPD), ratio of performance to interquartile range (RPIQ), and their arbitrary threshold values, to evaluate model capabilities and compare performance with other models [13], [15]. However, selecting the right performance metrics is crucial, as different measures highlight different aspects of model performance. Similar challenges arise across various fields including climate research, hydrology, machine learning, statistics, and spectroscopy, where relying solely on popular metrics can lead to different interpretations if used without proper context and complementary assessments [31], [32], [33], [34], [35]. Therefore, the authors [36], [37], [38] advocate that these metrics must be interpreted within the context of study's objectives, data characteristics, and value distributions. In large and heterogeneous SSLs, traditional measures may fail to capture model suitability, making additional metrics (e.g., MAE) and more detailed error analyses (e.g., error histograms, error correlation analysis) essential. Moreover, previous soil spectroscopy studies rarely assess structural independence i.e., verifying whether different models provide unique insights rather than replicating each other's predictions. Structural independence refers to the degree to which two models make different types of prediction errors and is assessed through error correlation analysis. Such analyses, routinely used in soil moisture and machine learning studies [39], [40], [41], can guide strategic model selection, and combination, potentially improving prediction in spectroscopy.

This study evaluates the application of NN and PLSR approaches for SOC prediction using the open source, large, diverse globally representative open soil spectral library (OSSL) [26], [27]. The extensive OSSL dataset provides a robust benchmark for comparing NN and PLSR models while evaluating the utility of various performance metrics. The primary objectives were to assess the accuracy of NN approaches in estimating SOC, compare their performance with PLSR models, and provide insights into SOC prediction across diverse conditions. Additionally, error correlation analysis was conducted to examine the structural independence of the models and ascertain the effect of performance metrics on model assessment. Further analyses were performed on selected models to investigate the effect of different calibration set size and sample splitting methods on model performance. The findings aim to inform future model selection, enhance SOC prediction strategies, and guide methodological best practices in soil spectroscopy.

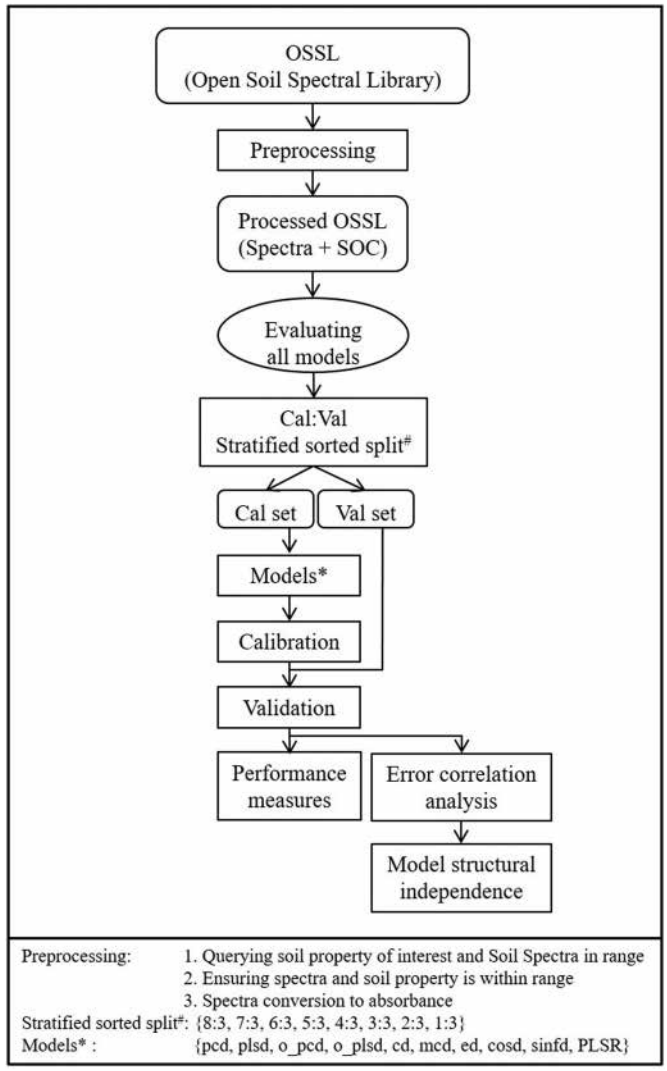


Fig. 1. Overall methodological framework, including soil spectral library preprocessing, dataset split, model calibration and evaluation.

II. Dataset, Models and Evaluation

A. Dataset

The OSSL[1] [26], [27] is a collaborative effort consisting of nine spectral libraries contributed by various organizations (details in Appendix section A). Samples were extracted from the OSSL by applying the following criteria (see Fig. 1):

1) the reported SOC content measured by the Walkley–Black method had to be $\geq 0.1\%$;
2) the VNIR spectra included had to range from 500 to 2500 nm, with a spectral resolution of 2 nm;

[1][Online]. Available: https://soilspectroscopy.org

TABLE I
DETAILS OF THE DISTANCE MEASURES USED BY EACH NN MODEL EMPLOYED IN THIS STUDY

Method	Dimensionality reduction	Distance measure	Formula
pcd	PCA	Mahalanobis	$md(i,j) = \sqrt{\frac{1}{n}\sum(p_i - p_j)C^{-1}(p_i - p_j)'}$
plsd	PLS	Mahalanobis	
o_pcd	PCA*	Mahalanobis	
o_plsd	PLS*	Mahalanobis	
cd	-	Correlation	$cd(i,j) = \frac{1-\rho(i,j)}{2}$
mcd	-	Moving correlation	$mcd(i,j) = \frac{1-\rho(i,j)_w}{2w}$
ed	-	Euclidean	$ed(i,j) = \sqrt{\frac{1}{n}\sum(i-j)(i-j)'}$
cosd	-	Cosine/Spectral angle mapper	$cosd(i,j) = cos^{-1}\left(\frac{i.j}{\|\|i\|\| * \|\|j\|\|}\right)$
sinfd	-	Spectral Information divergence	$sinfd(i,j) = kl(i,j) + kl(j,i)$

Note: All models are available in the resemble package in R (Ramirez-Lopez Et al., 2024 [48]).
*: optimal number of components was selected to minimize prediction error; (*i*,*j*): the distance between samples *i* and *j*; *n*: latent number of components; C^{-1}: inverse of covariance matrix computed from the matrix of projected variables; pi, pj : projected scores for samples *i* and *j* respectively; *w*: moving window size of 51 for spectra; kl: Kullback–Leibler divergence. *pcd* – principal component distance; *plsd* – partial least squares distance; *o_pcd* – optimized principal component distance; o_plsd – optimized partial least squares distance; *cd* – correlation distance; *mcd* – moving correlation distance; *ed* – Euclidean distance; *cosd* – cosine distance; *sinfd* – spectral information divergence.

3) both SOC values and reflectance spectra had to be within the theoretical ranges (0% < SOC content < 100%; 0 < Reflectance < 1).

The reflectance spectra were then converted to absorbance using (1) [42], [43]. This resulted in a dataset with 1000 spectral bands at 2 nm intervals in the VNIR range (500 to 2500 nm), each with a corresponding SOC content. The resulting refined SOC dataset (consisting absorbance spectra and their associated SOC values) was used for further analysis (see Fig. 1). After these refinements, a total of 63 321 samples with absorbance spectra and corresponding SOC content were obtained from the original OSSL dataset

$$\text{Absorbance} = -\log_{10}(\text{Reflectance}) \tag{1}$$

B. Models

SSLs provide enormous spectral data, but are subject to the "curse of dimensionality," [44], [45] where the high number of wavelengths per spectrum necessitates dimensionality reduction. Principal component analysis (PCA) and partial least squares (PLS) are standard decomposition methods in spectroscopic analysis [46], [47]. These methods assume that a subspace exists where the maximum amount of original information is retained. PCA identifies new orthogonal features that explain variance in the original spectral data, whereas PLS identifies features that explain variance in both the spectral data and the soil property of interest. The NN approach, part of nonlinear machine learning methods, is also known as instance or memory based learning, local modelling, cluster-based modeling or geographical segmentation-based modeling [48] while in different fields of study, it is also termed as NN allocation, NN imputation, One NN, KNN with $k = 1$, etc. PLSR predicts based on regression while NN identifies neighbors through clustering. Despite both methods using dimensionality reduction, PLSR and NN models differ fundamentally in their methodologies as one is a global regression approach while the other is an allocation based approach. In this study, a total of 10 models were evaluated: 1 PLSR model and 9 NN models (see Models in Fig. 1, Table I).

1) Partial Least Squares Regression Model: The PLSR model employs PLS decomposition followed by regression. In this study, the PLSR model used a ten-fold cross-validation with a one-sigma variation in RMSE minimization to determine the optimal number of components [49], [50], [51], [52]. The one-sigma variation selects the number of components within one standard error of the minimum RMSE, while the ten-fold cross-validation reduces computational time for model calibration (details in Appendix B and Fig. 4). In all models, the search for optimal number of components were limited to 100. Outliers were not removed in the analysis (both calibration and validation sets) as such values naturally occur, and the models must be robust enough to account for them [53], [54]. After calibration, the PLSR model applied weights to each wavelength of the absorbance spectra to predict SOC value for the validation set (see Fig. 1). No filtering or correction was applied to avoid biasing performance metrics. Negative SOC predictions produced by PLSR were retained in the analysis.

2) Nearest Neighbor Models: The NN models involve dimensionality reduction followed by similarity calculation, where the most similar neighbor is identified based on distance metrics. The proximity between two objects is inversely related to their distance, with shorter distances indicating greater similarity between samples [30]. The nine NN models evaluated in this study—*pcd, plsd, o_pcd, o_plsd, cd, mcd, ed, cosd,* and *sinfd*—each utilize distinct distance metrics for nearest neighbor identification (see Table I for a summary of the various distance measures, with abbreviations provided in the table footnote). The distance metrics were selected based on prior applications

in soil spectroscopy and chemometrics [48], [48]. The Cosine, Euclidean, Mahalanobis distance, etc., are commonly used for angular similarity in high-dimensional data. Specifically, the *pcd*, *plsd*, *o_pcd*, and *o_plsd* models apply Mahalanobis distance within a latent space, whereas *cd* and *mcd* rely on correlation-based measures, with *mcd* incorporating a moving window of size 51. The *ed* model employs Euclidean distance, *cosd* utilizes Cosine distance, and *sinfd* is based on spectral information divergence [48]. For the *pcd* and *plsd* models, the number of components (principal components for *pcd* and latent/pls components for *plsd*) was selected based on variance explained. For the *o_pcd* and *o_plsd* models, the optimal number of components was determined by RMSE minimization using leave-one-out cross-validation (see Fig. 1).

After calibration, the NN models computed the distance between each validation sample and the calibration sample. The distance measures and the number of components varied depending on the method (see Table I). For each validation sample, the nearest neighbor was selected based on the minimum distance. The predicted SOC value for sample in calibration set was the SOC value of the selected neighbor in calibration set (see Fig. 1). This process was repeated for all validation samples to predict their SOC values.

C. Model Evaluation

To evaluate model quality, the SOC dataset was split into calibration and validation sets using stratified sorted sampling based on SOC values [55], [56], [57] [see Fig. 1(b) and (c)]. This sampling approach ensured similar distributions and coverage of full SOC value range across both sets (details in Table IV of Appendix A). The calibration:validation ratio was set to 8:3 for model evaluation [see Fig. 8(b) of Appendix E]. Specifically, samples were sorted in increasing order of SOC values. To ensure similar distribution and SOC range in both sets, the first eight samples were assigned to the calibration set and the next three to the validation set. This process was repeated until all samples were categorized. To ensure a fair comparison across the 10 models (9 NN and 1 PLSR), this stratified sorted sampling was applied consistently (see Fig. 1) i.e., the models were trained on the same calibration set and were evaluated on the same validation set. The total dataset of 63 321 samples was split into 46 053 calibration samples and 17 268 validation samples. After calibrating the models on the calibration set, predictions were made for both the calibration and validation sets. This section details how the framework was evaluated through the following three main analyses:

1) using traditional performance measures;
2) detailed error analysis and testing structural independence via error correlation analysis;
3) assessing effect of calibration set size and sampling method.

1) Traditional Performance Measures: Each model was evaluated by comparing predicted SOC values with observed SOC values in both the calibration and validation sets. The traditional metrics used were mean absolute error (MAE), R^2, RMSE, RPD, RPIQ, and bias (2)–(7) [31], [38], [54], [57], [58]. These measures are routinely used for model assessment in soil spectroscopy. The RPIQ metric is a complementary metric to RPD, especially useful when the data distribution is skewed. It is computed as the ratio of the interquartile range of observed values to the RMSE of predictions. Higher RPIQ values indicate better model performance, and it can offer more robustness than RPD in datasets with outliers or non-normal distributions [38]. Based on the values of $\underline{R}^2$ and RPD, [59] classify models into three categories with decreasing confidence in the predictions: good predictions ($R^2 > 0.8$; RPD > 2), moderate predictions ($0.5 < R^2 < 0.8$; $1.4 <$ RPD < 2) or poor predictions ($R^2 < 0.5$; RPD < 1.4)

$$\text{MAE} = \frac{\sum \left| Y - \hat{Y} \right|}{n} \tag{2}$$

$$R^2 = 1 - \left[\frac{\sum \left(Y - \hat{Y} \right)^2}{\sum \left(Y - \bar{Y} \right)^2} \right] \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum \left(Y - \hat{Y} \right)^2}{n}} \tag{4}$$

$$\text{RPD} = \frac{\text{SD}}{\text{RMSE}} \tag{5}$$

$$\text{RPIQ} = \frac{\text{IQR}}{\text{RMSE}} \tag{6}$$

$$\text{bias} = \bar{\hat{Y}} - \bar{Y} \tag{7}$$

In which, Y correspond to observations; $\hat{Y}$ to model predictions; $\bar{Y}$ is mean, SD is standard deviation and IQR is interquartile range from the observations; $\bar{\hat{Y}}$ is mean from the model predictions; and n is the number of samples.

2) Error Analysis: Detailed error analysis (error histograms and error correlation analysis) provides insights into the model's performance and allows for cross-checking inferences from traditional performance measures. The error histograms capture each model's distribution of errors providing insights on model's behavior, while error correlation analysis helps confirm whether two models are structurally independent [39], [40], [41]. Since many methods use similar distance measures (e.g., Mahalanobis, Euclidean), transformed space (PLS, PCA) and component selection frameworks (see Table I, Fig. 1), it is crucial to verify the structural independence of the models i.e., the models are different from each other and not merely replicas. Error correlation analysis helps in testing this structural independence. A high correlation of errors between two models indicates structural dependence, while a low or near-zero correlation suggests independence. General interpretation of error correlation includes: positive correlations indicate models make similar mistakes while negative correlations indicate that the models tend to make mistakes in opposite directions. Near-zero correlations indicate that errors are independent. Thus, to ensure appropriate model evaluation, first traditional performance measures are quantified followed by model structural independence, finally inferences are cross-checked by error histogram analysis.

3) Evaluating Effect of Calibration Set Size and Sampling Method: Finally, the effect of calibration set sizes and different sample splitting methods, on model performance was examined. To assess the effect of calibration set sizes, the SOC dataset was split into calibration and validation sets using the ratios varying from 8:3 down to 1:3 [see Fig. 8(b) of Appendix E]. These ratios were chosen to reflect a range of conditions from data-rich to data-sparse scenarios. Similar proportions have been tested in earlier studies assessing calibration size sensitivity in soil spectroscopy [60], [61], [62], [63], [64]. For each ratio, the model performance was evaluated on a fixed validation set while the calibration set size was varied, ensuring similar SOC value distributions in both sets. Specifically, the evaluated split ratios included 8:3, 7:3, 6:3, 5:3, 4:3 3:3, 2:3, and 1:3 (see Table IV of Appendix A; Fig. 8 of Appendix E).

To assess the effect of sample split, two sample splitting methods were used: stratified sorted split and random split. Models were evaluated over 100 bootstrap iterations to assess the effect of random sampling [for evaluating sampling method variations; Fig. 8(c) of Appendix E]. Mean and standard deviation for each model and sampling method were tabulated. These mean performance measures from both sampling methods and models were used to calculate the relative change in percentage for split size of 8:3 and 3:3. This analysis also enabled testing the effect of calibration set size variations and the choice of sample splitting methods on model performances [see Figs. 1; 8(b) and 8(c) of Appendix E).

All analyses were carried out using R (version 4.3.3; [65]) with the "ggplot2" package for plotting, the "resemble" package for NN models [48] and the "pls" package for PLSR model implementation. Computations were executed on an Intel Xeon 5320 CPU system (2.20 GHz processor, 512 GB RAM, NVIDIA GeForce RTX 4060Ti 16 GB graphics card). The average computational time for a complete set of analyses was approximately 10 days. Figures were prepared using OriginPro 2024b. The analysis-ready data and the code utilized for reproducing the results in this paper can be found on GitHub[2] and Zenodo [66].

III. Results

A. Exploratory Analysis of Spectra and SOC Values

The characteristics of raw absorbance spectra and the SOC statistics for all 63 321 samples are shown in Fig. 2(a). The typical spectral behavior of soils is observed [see Fig. 2(b)], with distinct absorption peaks around 1400 and 1900 nm, attributed to water, and near 2200 nm, associated with clay minerals [10]. An increase in SOC content corresponds to higher absorbance values, and stratification of spectra by SOC quartiles [see Fig. 2(b)] demonstrates a systematic rise in overall absorbance with increasing SOC levels. Notably, samples with very high SOC content exhibit a reduction in the clay-related absorption peak [60], [61], [62], [63], [64]. The complete range of soil absorbance variability across all samples is provided in the Supplementary Material (see Fig. S1).

[2][Online]. Available: https://github.com/ternikarcr/SOC_NN.git

The PCA of the spectral data [see Fig. 2(a)] shows that the first three principal components (PCs) explain 98.77% of the spectral variance, while the first ten PCs explain nearly all of the variance (99.96%). The SOC content ranges from 0.1% to 78.45%, with a mean of 6.77%, a median of 1.91%, standard deviation of 12.50%, and an interquartile range of 3.42% [see Fig. 2(a)]. The mean value exceeding the median, the standard deviation surpassing the interquartile range and a skewness value of 2.6 collectively indicate that SOC data is highly skewed [see Fig. 2(a)].

B. Performance of NN and PLSR Models

All models demonstrated varying degrees of accuracy in predicting SOC, with performance ranging from good to moderate. The calibration and validation results were generally consistent, suggesting that models neither overfit nor underfit (see Table II(a); Fig. 5 in Appendix C). For the validation set, the MAE ranged from 1.79% to 4.22% [see Fig. 2(e)], the RMSE from 3.75% to 9.23%, and R^2 from 0.53 to 0.91 (see Table II(a)). Among the NN models, *o_plsd* exhibited the highest performance (MAE = 1.79%, RMSE = 3.88%, R^2 = 0.91), closely followed by *o_pcad* (MAE = 1.82%, RMSE = 3.93%, R^2 = 0.90). Other well-performing NN models included *mcd* and *ed*, while *sinfd* was the weakest model (MAE = 4.22%, RMSE = 9.23%, R^2 = 0.53), (see Table II(a)). The PLSR model also performed well (MAE = 2.36%, RMSE = 3.75%, R^2 = 0.91), aligning with the best-performing NN models. Based on classification thresholds from [59], five models (*o_plsd, o_pcad, mcd, ed,* and PLSR) provided good predictions, while the remaining five (*pcad, plsd, cd, cosd, sinfd*) delivered moderate performance.

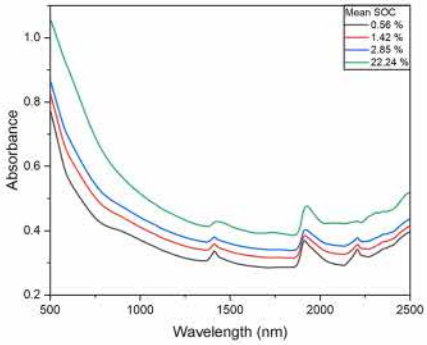
Notably, the choice of the best model depends on the selected performance metric. For example, based solely on RMSE, PLSR appears to perform best (3.75%), slightly outperforming *o_plsd* (3.88%), *o_pcad* (3.93%), and *mcd* (5.17%). However, when considering MAE, *o_plsd* ranks highest (1.79%), followed by *o_pcad* (1.82%), *mcd* (2.35%), and PLSR (2.36%). In terms of explained variance (R^2), both *o_plsd* and PLSR achieve the highest value (R^2 = 0.91), followed by *o_pcad* (0.90) and *mcd* (0.84). When considering all metrics collectively, *o_plsd, o_pcad,* and PLSR emerge as the top three models. This variation in model ranking depending on the metrics underscores the importance of conducting a detailed error analysis, especially when dealing with large datasets with diverse value distributions. Relying solely on traditional performance measures may be insufficient to identify the most effective model.

C. Error Analysis of NN and PLSR Models

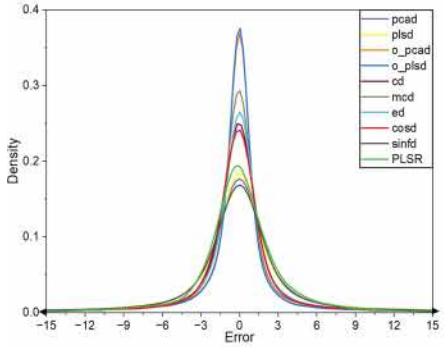
To understand the discrepancies in selecting the best model based on performance measures, as well as to test the structural independence of the models, a detailed error analysis (error histograms, error correlation analysis) was conducted. The error correlation analysis of the validation set identifies structural similarities between two pairs of models: (*pcad, plsd*) and (*o_pcad, o_plsd*), with correlation coefficient values of 0.58 and 0.66, respectively [see Fig. 2(f)] [67]. Their high error correlations

OSSL (63321 samples)			
Soil Spectra (Absorbance)		Soil property (SOC %)	
PCs	Var Explained (%)	Statistics	Value (%)
1	87.03	Min	0.10
2	9.42	Max	78.45
3	2.32	Mean	6.77
4	0.58	SD	12.50
5	0.37	Median	1.91
6	0.13	IQR	3.42
7	0.05	CV	1.85
8	0.03	Skewness	2.60
9	0.02	Kurtosis	8.65
10	0.01	Q1	1.01
Total	99.96	Q3	4.43

(a)



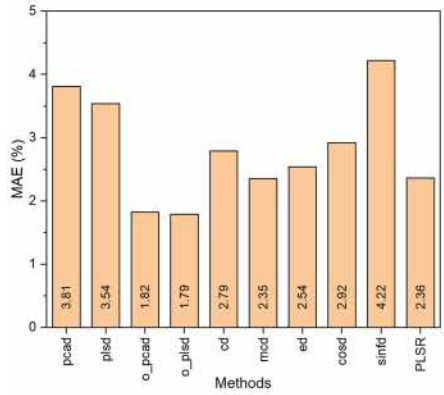
(b)



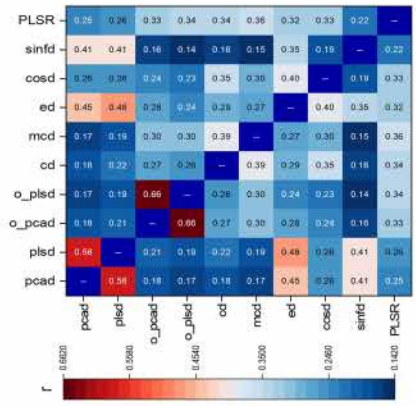
(c)

Error Statistics						
Methods	Min	Max	Mean	SD	Median	IQR
pcad	-55.46	55.47	0.04	7.85	0.01	2.69
plsd	-59.51	58.74	0.02	7.18	0.01	2.68
o_pcad	-45.63	46.11	-0.18	3.93	-0.02	1.26
o_plsd	-46.48	46.11	-0.17	**3.88**	-0.02	**1.21**
cd	-54.04	61.03	-0.05	6.36	-0.03	1.65
mcd	-48.08	48.66	0.01	5.17	-0.04	1.51
ed	-56.21	50.16	-0.07	5.35	0.01	1.80
cosd	-54.86	58.01	0.38	6.68	-0.03	1.70
sinfd	-63.31	56.34	0.05	9.23	0.01	2.55
PLSR	-43.99	31.56	0.02	**3.75**	0.00	**2.87**

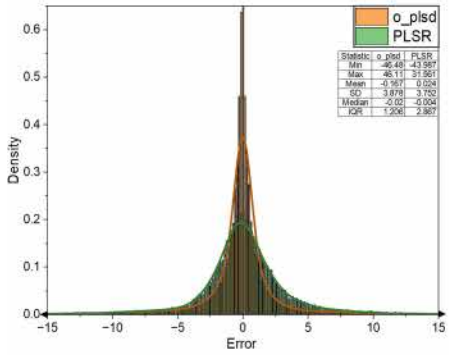
(d)



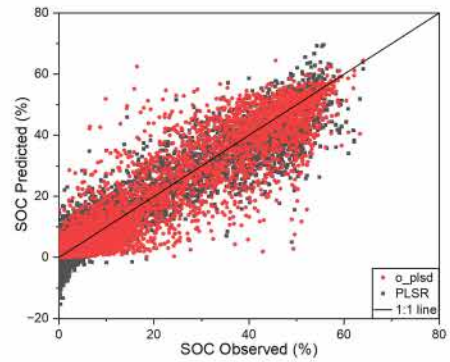
(e)



(f)



(g)



(h)

Fig. 2. (a) Variance explained by principal component spectra and summary statistics of SOC values from the processed OSSL. (b) Mean VNIR spectra stratified by SOC quartiles. (c) Overlaid error density plots for all ten models in the validation set using the stratified sorted 8:3 calibration:validation split. (d) Error statistics table summarizing model performance. (e) Variation in MAE across all models in the validation set. (f) Error correlation matrix for all model pairs in the validation set. (g) Overlaid error density plots comparing *o_plsd* and PLSR. (h) Predicted versus observed scatterplots for *o_plsd* and PLSR in the validation set.

TABLE II
PERFORMANCE EVALUATION SUMMARY FOR BOTH CALIBRATION AND VALIDATION SETS, INCLUDING. (A) PERFORMANCE OF ALL MODELS; (B) PERFORMANCE VARIATIONS DUE TO INPUT (CALIBRATION SET) SIZE CHANGES, AND (C) PERFORMANCE VARIATIONS DUE TO DIFFERENT SAMPLING METHODS

Sampling method	Methods	Split ratio	Split ratio (%)	Split Nos	Calibration							Validation					
					Ncomp	MAE	R^2	RMSE	RPD	RPIQ	Bias	MAE	R^2	RMSE	RPD	RPIQ	Bias
(a) All models																	
Stratified	*pcad*	8:3	72.7:27.3	46053:17268	3	3.80	0.64	7.86	1.59	0.44	–0.03	3.81	0.65	7.85	1.59	0.43	–0.04
Stratified	*plsd*	8:3	72.7:27.3	46053:17268	3	3.63	0.69	7.34	1.70	0.47	–0.01	3.54	0.70	7.18	1.74	0.48	–0.02
Stratified	*o_pcad*	8:3	72.7:27.3	46053:17268	25	1.78	0.91	3.81	3.28	0.90	0.21	1.82	0.90	3.93	3.18	0.87	0.18
Stratified	*o_plsd*	8:3	72.7:27.3	46053:17268	24	1.75	0.91	3.78	3.31	0.90	0.22	**1.79**	**0.91**	**3.88**	**3.22**	**0.88**	0.17
Stratified	*cd*	8:3	72.7:27.3	46053:17268	NA	2.76	0.77	6.28	1.99	0.54	–0.06	2.79	0.76	6.36	1.97	0.54	0.05
Stratified	*mcd*	8:3	72.7:27.3	46053:17268	NA	2.37	0.83	5.32	2.35	0.64	–0.01	2.35	0.84	5.17	2.42	0.66	–0.01
Stratified	*ed*	8:3	72.7:27.3	46053:17268	NA	2.53	0.82	5.36	2.33	0.64	0.05	2.54	0.82	5.35	2.34	0.64	0.07
Stratified	*cosd*	8:3	72.7:27.3	46053:17268	NA	2.90	0.76	6.61	1.89	0.52	–0.37	2.92	0.75	6.69	1.87	0.51	–0.38
Stratified	*sinfd*	8:3	72.7:27.3	46053:17268	NA	4.19	0.54	9.16	1.37	0.37	–0.11	4.22	0.53	9.23	1.35	0.37	–0.05
Stratified	*PLSR*	8:3	72.7:27.3	46053:17268	73	2.31	0.92	3.63	3.45	0.94	0.00	**2.36**	**0.91**	**3.75**	**3.33**	**0.91**	–0.03
(b) Variation in input/calibration dataset																	
Stratified	*o_plsd*	8:3	72.7:27.3	46053:17268	24	1.75	0.91	3.78	3.31	0.90	0.22	**1.79**	0.91	3.88	3.22	0.88	0.17
Stratified	*o_plsd*	7:3	70.0:30.0	40297:17268	23	1.76	0.91	3.80	3.29	0.90	0.21	1.78	0.91	3.83	3.27	0.89	0.20
Stratified	*o_plsd*	6:3	66.6:33.4	34541:17268	17	1.80	0.91	3.85	3.25	0.89	0.19	1.81	0.91	3.83	3.26	0.89	0.17
Stratified	*o_plsd*	5:3	62.5:37.5	28785:17268	27	1.85	0.90	3.95	3.17	0.87	0.26	1.88	0.90	4.08	3.07	0.84	0.21
Stratified	*o_plsd*	4:3	57.1:42.9	23028:17268	14	1.91	0.90	4.02	3.11	0.85	0.23	1.92	0.90	4.00	3.13	0.85	0.19
Stratified	*o_plsd*	3:3	50.0:50.0	17271:17268	27	1.94	0.89	4.11	3.04	0.83	0.28	1.97	0.89	4.18	2.99	0.82	0.23
Stratified	*o_plsd*	2:3	40.0:60.0	11514:17268	23	2.01	0.89	4.19	2.99	0.82	0.29	2.04	0.89	4.24	2.95	0.80	0.21
Stratified	*o_plsd*	1:3	25.0:75.0	5757:17268	26	2.12	0.88	4.43	2.83	0.77	0.40	**2.23**	0.86	4.74	2.64	0.72	0.31
Stratified	PLSR	8:3	72.7:27.3	46053:17268	73	2.31	0.92	3.63	3.45	0.94	0.00	**2.36**	0.91	3.75	3.33	0.91	–0.03
Stratified	PLSR	7:3	70.0:30.0	40297:17268	71	2.31	0.92	3.63	3.44	0.94	0.00	2.37	0.91	3.76	3.33	0.91	–0.02
Stratified	PLSR	6:3	66.6:33.4	34541:17268	66	2.32	0.91	3.66	3.42	0.93	0.00	2.38	0.91	3.78	3.31	0.90	–0.02
Stratified	PLSR	5:3	62.5:37.5	28785:17268	64	2.32	0.91	3.66	3.42	0.93	0.00	2.39	0.91	3.79	3.30	0.90	–0.04
Stratified	PLSR	4:3	57.1:42.9	23028:17268	57	2.34	0.91	3.69	3.39	0.93	0.00	2.41	0.91	3.83	3.27	0.89	–0.04
Stratified	PLSR	3:3	50.0:50.0	17271:17268	56	2.34	0.91	3.68	3.40	0.93	0.00	2.42	0.91	3.85	3.25	0.89	–0.05
Stratified	PLSR	2:3	40.0:60.0	11514:17268	50	2.37	0.91	3.67	3.41	0.93	0.00	2.45	0.90	3.87	3.23	0.88	–0.05
Stratified	PLSR	1:3	25.0:75.0	5757:17268	39	2.47	0.91	3.78	3.31	0.90	0.00	**2.56**	0.90	4.02	3.11	0.85	–0.11
(c) Variation in sampling method																	
Random*	*o_plsd*	8:3	72.7:27.3	46053:17268	20	1.75	0.91	3.77	3.29	0.90	0.18	**1.80**	0.91	3.90	3.28	0.88	0.18
Random*	PLSR	8:3	72.7:27.3	46053:17268	71	2.31	0.91	3.64	3.40	0.94	0.00	**2.36**	0.91	3.74	3.42	0.92	0.05
Random*	*o_plsd*	3:3	50.0:50.0	17271:17268	21	2.21	0.89	4.52	3.06	1.00	0.27	**1.96**	0.89	4.12	3.04	0.83	0.16
Random*	PLSR	3:3	50.0:50.0	17271:17268	54	2.58	0.92	3.97	3.48	1.14	0.00	**2.49**	0.91	3.86	3.24	0.88	–0.09

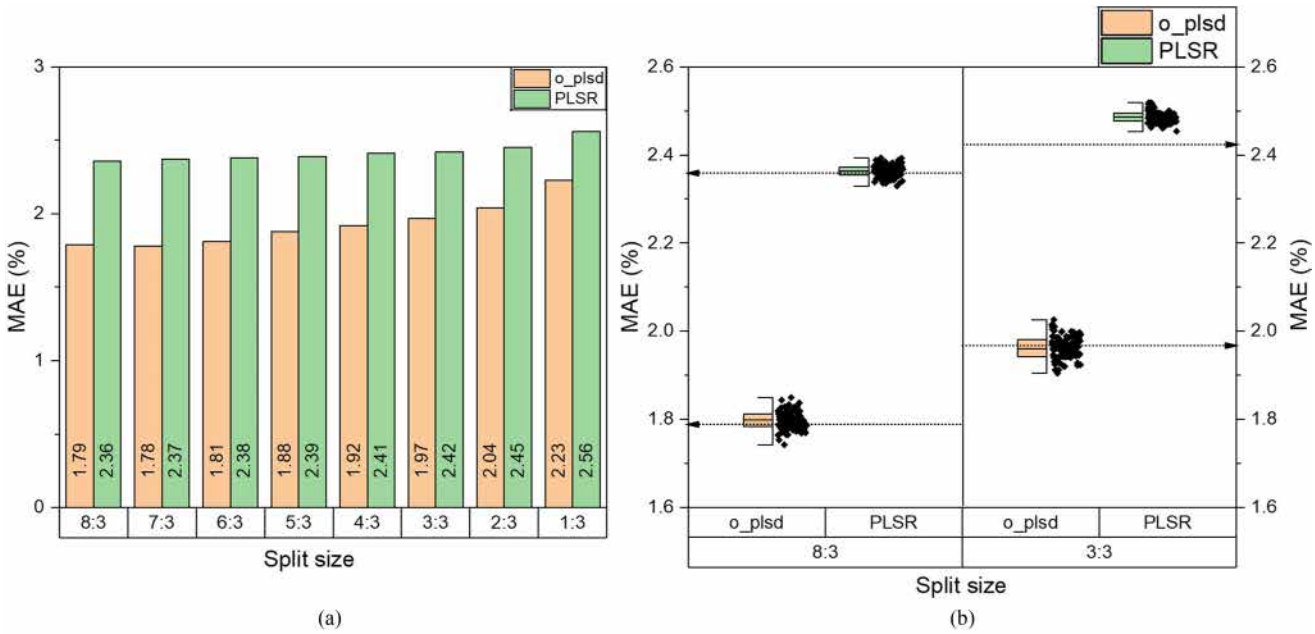
Note: *Bootstrap 100 iterations with mean values reported in table.

indicate that *o_plsd* and *o_pcad* are structurally similar, meaning that using both does not provide additional insights into SOC predictions beyond what one model can offer. Although these pairs utilize different dimensionality reduction techniques (PCA versus PLS; see Table I), they are still structurally similar and thus do not contribute distinct information. Other model pairs exhibited low error correlations, suggesting structural independence. For instance, the error correlation between *o_plsd* and PLSR (both using PLS based dimensionality reduction) is relatively low [0.34, see Fig. 2(f)], indicating that these two models provide independent insights into SOC predictions [see Fig. 2(e)]. In contrast, other model pairs remain independent. Thus, among the top three models identified by performance measures (*o_plsd*, *o_pcad*, PLSR), only o_plsd and PLSR are structurally independent. Additional insights are provided in Appendix section C.

To compare the performance of *o_plsd* and PLSR, error histograms were analyzed. The *o_plsd* model exhibited a higher concentration of errors near zero and a more compact error distribution, whereas PLSR displayed a broader error spread with a longer tail [see Fig. 2(c) and (d)]. Overlaid scatter plots [see Fig. 6(b) and (c) in Appendix Section C] and error histograms [see Fig. 2(c)] further indicate that *o_plsd* predictions are more tightly clustered around the 1:1 line compared to PLSR, which even produced some negative SOC predictions [see Fig. 2(g) and (h)]. Although *o_plsd* has a slightly wider overall error range (–46.5% to 46.1%) compared to PLSR (–43.9% to 31.6%), its interquartile range (IQR) is notably smaller (1.2% for *o_plsd* versus 2.9% for PLSR) [see Fig. 2(d)]. Additionally, the errors in *o_plsd* are more densely concentrated near zero, reinforcing its superior predictive accuracy. Considering both the structural independence of the models and the characteristics of the error distribution, *o_plsd* emerges as the best-performing model, followed by PLSR [see Fig. 2(d), (g), and (h); Table II(a)].

D. Evaluating Effect of Calibration Set Size and Sampling Method on O_Plsd and PLSR Models

Two sets of analyses were conducted to evaluate the effect of calibration set size and sampling methods on model performance of *o_plsd* and PLSR. First, the effect of varying the calibration set size was examined (see Fig. 1). Second, the influence of using different sample splitting methods (stratified versus random) was tested. Finally, to quantify the actual contributions from calibration set size and sample splits, relative changes were computed. Because MAE was consistent with the detailed error analysis and more appropriate for intercomparison studies [34], it is used to summarize the results.



o_plsd

Split size | Sampling	Stratified	Random	relative Δ (%)
8:3	1.79	1.80	0.55
3:3	1.97	1.96	-0.51
relative Δ (%)	10	8.88	

(c)

PLSR

Split size | Sampling	Stratified	Random	relative Δ (%)
8:3	2.36	2.36	0.00
3:3	2.42	2.49	2.42
relative Δ (%)	2.54	5.51	

(d)

Fig. 3. Effect of calibration set size and sampling methods on model performance in validation set. (a) Barplot showing variation in MAE for *o_plsd* and PLSR as calibration:validation ratio decreases from 8:3 to 1:3. (b) Boxplots illustrating MAE distributions, from 100 bootstrap random splits at 8:3 and 3:3 ratios, with dotted lines indicating stratified splits for *o_plsd* and PLSR. (c) Tabulated mean MAE and relative changes due to calibration set size and sampling method choice for the *o_plsd* model (d) Same as (c) but for the PLSR model.

1) Effect of Calibration Set Size: As expected, reducing the calibration set size resulted in increased errors for both the *o_plsd* and PLSR models. For *o_plsd*, the MAE increased from 1.79% (at an 8:3 calibration:validation split) to 2.23% (at a 1:3 split), while for PLSR it rose from 2.36% to 2.56% (see Table II(b); Figs. 3(a); 9 of Appendix E). Although *o_plsd* consistently outperformed PLSR at all calibration:validation ratios, its MAE exhibited a steeper increase, indicating a higher sensitivity to calibration set size. In contrast, PLSR showed a more gradual increase in MAE, suggesting greater generalizability (Appendix Section E). Notably, the MAE increase for *o_plsd* (2.23–1.79 = 0.44%) was twice that of PLSR (2.56–2.36 = 0.20%).

2) Effect of Sampling Method: Next, 100 bootstrap iterations were performed using random calibration-validation splits to assess the effect of sampling methods on both models [see Fig. 3(b)]. Both *o_plsd* and PLSR show similar performance when the calibration and validation sets are selected randomly versus using stratified sorting. For instance, *o_plsd* has a validation MAE of 1.79% with stratified and 1.80% with random sampling at the 8:3 ratio—a negligible difference. Similarly, PLSR had no difference with MAE of 2.36% both in random and stratified sampling (see Table II(c)). Although *o_plsd* consistently outperformed PLSR, its MAE fluctuated about twice as much as PLSR's under random sampling iterations. Nonetheless, these variations are much smaller than those observed when reducing calibration set size.

3) Relative Contributions of Calibration Set Size and Sampling Method: To quantify the actual contributions from calibration set size and sample splits, relative changes were computed. For the 8:3 split, *o_plsd* had an average MAE of 1.79% with stratified sampling and 1.80% with random sampling—a relative change of about 0.55%. Similar relative changes were computed for both models, different calibration set sizes, and sampling methods [see Fig. 3(c) and (d)]. For *o_plsd*, the relative change in MAE due to calibration set size variation was 8.88% –10%, whereas the relative change due to sampling method was only –0.51% –0.55% [see Fig. 3(c)]. Thus, the effect of calibration set size on *o_plsd's* MAE is about ten times greater than that of the sampling method. For PLSR, the relative change in MAE due to calibration set size variation was 2.54% –5.51%, while the change attributable to sampling method choice was 0% –2.42% [see Fig. 3(d)]. Thus, calibration set size had roughly twice the influence of sampling method on PLSR's performance. Comparing the two models, *o_plsd* is approximately twice as affected by calibration set size variation and about half as affected by sampling method choice relative to PLSR.

In summary, *o_plsd* required fewer components (24) than PLSR (73) and consistently showed lower MAE values (1.79% versus 2.36%), achieving higher accuracy, suggesting efficiency in handling spectral data. Although *o_plsd* was more accurate than PLSR, reducing the calibration data disproportionately affected its performance, causing faster degradation as the calibration set reduced. By contrast, the choice between stratified and random sampling has a relatively minor effect on performance; while *o_plsd* was less affected, PLSR showed slightly greater sensitivity in this regard. Overall, these findings for SOC prediction suggest that *o_plsd* is well suited for tasks requiring high accuracy when ample calibration data is available, whereas PLSR may be preferable when generalizability is needed and the calibration set is very limited.

IV. Discussion

A. Effectiveness of O_Plsd and PLSR Models

The *o_plsd* model achieved a low MAE of 1.79%, RMSE of 3.88%, and an RPD of 3.22, while PLSR attained a MAE of 2.36%, RMSE of 3.75%, and an RPD of 3.33. Both these RPD values fall within the category of "excellent models" [68] or "good predictions" [59], illustrating the effectiveness of both VNIR spectroscopy and o_plsd and PLSR models for predicting SOC in diverse global soils. The performance metrics of the models in this study align well with findings from meta-analyses and individual studies on VNIR spectroscopy for SOC prediction. Meta-analyses by Ahmadi et al. [13], and Chinilin et al. [15] reported mean and median R^2 values of 0.76 and 0.67, respectively, while the models in this study exceed those values. However, the RMSE here is higher, likely due to the large sample size (63 321) and the broad SOC range (0.1% –78.45%), which naturally contribute to increased errors despite strong model performance. Comparisons with individual studies further highlight the robustness of the models evaluated here. For instance, [27], using the OSSL with a Cubist model, achieved $R^2 = 0.94$, RMSE = 2.36%, slightly outperforming this study but remaining broadly comparable. Similarly, study [69], analyzing 19 804 U.S. soils [70], achieved $R^2 = 0.83$, RMSE = 7.38% using PLSR and $R^2 = 0.96$, RMSE = 3.61% with ANN. National and regional soil spectral libraries often report high accuracy but generally involve smaller datasets and narrower SOC ranges, leading to lower RMSE values. For example, Brown et al. [71] reported $R^2 = 0.87$, RMSE = 0.4%, and [37] found RMSE = 0.25% in Australian soils. Studies from Europe, China, Brazil, and India [54], [72], [73], [74] also demonstrate strong results, but with more localized datasets, such as study [73] achieving $R^2 = 0.90$, RMSE = 0.37% for Chinese soils. While some studies report lower RMSE values, differences in dataset size, SOC range, and modeling approaches must be considered. Given the large, globally distributed dataset and the exceptionally wide SOC range in this study, the models perform competitively compared to existing literature.

B. Choosing Performance Measures Carefully

In large datasets with wide range of values, traditional metrics alone may not fully capture predictive accuracy or model suitability. In this study, a combination of error analysis, R^2, and MAE provided a more nuanced understanding of model performance than any single metric. For example, based solely on RMSE, PLSR performs best and when considering MAE, *o_plsd* performs best while in terms of R^2, both *o_plsd* and PLSR achieve the highest value. This highlights the importance of carefully selecting performance measures when dealing with large sample sizes and wide range of values. Furthermore, performance measures alone can be misleading if models are structurally similar. Error correlation analysis should be conducted to determine whether models offer independent insights or simply replicate one another's predictions. Such structural assessments may guide the combination of models, potentially improving SOC prediction accuracy. Finally, to ensure transparency and fair comparisons, future research should consider providing pairs of predicted and observed SOC values. As advocated by Shao et al. [75], sharing these data enables the broader research community to apply different performance metrics and methods of structural assessment, ensuring consistent, reproducible evaluations and guiding the development of more informative performance measures.

Selecting the right performance metrics is critical, as different measures highlight different aspects of model performance. Multiple classification schemes have been proposed for interpreting the RPD metric in soil spectroscopy: for example, Chang et al. [59] defined three classes of model performance with RPD thresholds of 1.4 and 2.0, whereas Chinilin et al. [15], and Rossel et al. [76] proposed three class systems with slightly different cutoffs (1.5 and 2.0), and Saeys et al. [68], and Shi et al. [73] offered more detailed categorizations with thresholds at 1.5, 2.0, 2.5, and 3.0 to distinguish between poor, approximate, moderate, good, and excellent predictions. Similar thresholds also apply for R^2 and other performance metrics. Although these cutoffs are widely used in soil spectroscopy, several authors caution against over-reliance on fixed RPD thresholds [36], [37]. Minasny and McBratney [38] emphasize that performance measures should reflect the study's objectives and data characteristics rather than arbitrary thresholds. Such caution extends beyond spectroscopy to metrics like KGE and NSE in hydrology [32], kurtosis in statistics [70], overall versus average accuracy or Matthews correlation coefficient versus F1 in machine learning [35], [75], [77], [78] and RMSE versus MAE in climate studies [34]. All these highlight how different metrics can lead to different interpretations if used without proper context and complementary assessments.

C. Advantages of O_Plsd and PLSR Models

The superior performance of the *o_plsd* model over PLSR in this study can be attributed to both data characteristics and methodological differences. The *o_plsd* model operates within the range of the input data and capitalizes on the density and proximity of similar samples, ensuring predictions remain

realistic and well-anchored within the observed sample space. By leveraging local variations in the data, *o_plsd* enhances accuracy, particularly in cases where samples are densely clustered around specific SOC levels. In contrast, PLSR, as a global regression-based approach, does not inherently consider local density or proximity. It can produce predictions outside the observed data range, particularly in regions with sparse sample coverage. PLSR assumes a primarily linear relationship and may not capture the complexity or nonlinearity of the spectra-SOC relationship, resulting in limited performance when wide range of values are present. Hence, while PLSR generalizes well with smaller datasets, it is less sensitive to subtle, localized patterns that strongly influence *o_plsd's* accuracy (Table V in Appendix D). This explains why *o_plsd* provides better performance, particularly for a densely populated OSSL dataset. Previously, Ng et al. [79] indicated that PLSR provides good predictions, often outperforming more complex models, especially under varying sampling conditions. This robustness explains why PLSR in this study exhibits lower sensitivity to calibration set size variations. It remains comparatively stable when data are limited or unevenly distributed. However, when a rich, densely populated dataset is available, *o_plsd* outperforms PLSR by better leveraging local spectral information and improves SOC predictions.

D. Effect of Calibration Set Size

In this study, reducing the calibration set size led to a varying degree of increased predictive errors in both models. The MAE decreased for *o_plsd* (2.23% to 1.79%) and PLSR (2.56% to 2.36%) when calibration size increased from ∼9% to 72.7% of total sample size. This is consistent with numerous studies that show improved performance in VNIR-based SOC prediction with increasing number of calibration samples [12], [22], [71], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88]. This improvement is generally attributed to the more comprehensive representation of soil variability viable by a larger calibration set. For example, Shepherd and Walsh [80] reported an increase in R^2 from ∼ 0.20 to 0.75 when the calibration set proportion rose from 5% to 67% of the total samples. Similarly, Kuang and Mouazen [81] observed that error decreased almost linearly (from 0.78% to 0.64%) as the calibration set grew from ∼20% to 80% of total samples. In another instance, Grinand et al. [82] showed error reductions from 0.80% to 0.59% when the calibration size ranged from 10% to 80% of total samples, and Clairotte et al. [86] similarly noted re ductions from 0.59% to 0.48% as the calibration increased from 25% to 90%. Collectively, these findings underscore the importance of an adequately sized calibration set in capturing soil complexity and enhancing model performance.

E. Effect of Sampling Method

All sampling methods rely on different principles for selecting calibration samples. However, except the random sampling, others employ a stratification strategy to enhance sample representativeness. In this study, the choice of sampling method did not significantly affect the calibration performance, as both models exhibited negligible changes in MAE likely due to the large overall sample size (validation set of 17 268). This is consistent with previous studies demonstrating minimal or negligible impact on performance in VNIR-based SOC prediction with varying sampling methods [62], [79], [83], [85], [88], [89]. Such minimal variation is generally attributed to a sufficiently broad representation of soil variability within both calibration and validation sets [88]. On the contrary, several studies have demonstrated improved performance using stratified sampling methods over random sampling—particularly for smaller datasets. For example, Clairotte et al. [86] employed Kennard-Stone sampling on a validation set of 380 samples, Debaene et al. [83] using *k*-means clustering with 199 samples and Ng et al. [79] applied conditioned Latin hypercube sampling with 1000 samples, all noting better results compared to random sampling. Ramirez-Lopez et al. [85] similarly emphasized that the choice of sampling method is important when the calibration set is relatively small, whereas it is less consequential for larger datasets. Collectively, these findings underscore that sufficiently capturing soil variability in both calibration and validation sets is vital. Thus, random sampling in large datasets may be adequate, provided variability is comprehensively captured. In this aspect, Brown et al. [89], and Soriano-Disla et al. [90] advocate choosing calibration samples that represent the population's variability, match the intended model use and reflect spatial structure, thereby ensuring reliable predictions for independent validations.

F. Limitations

While this study provides robust benchmarking of NN and PLSR models using a large, global spectral dataset, several limitations should be noted. First, the analysis is based solely on the OSSL dataset; model generalizability to region-specific SSLs or new field-acquired spectra remains untested. Second, while *o_plsd* achieves high accuracy, it is computationally expensive and less interpretable than linear models. The complete *o_plsd* calibration and validation cycle required ∼32 h, whereas PLSR completed it in 12 h. The higher complexity of *o_plsd* arises from distance computations for all calibration-validation pairs. These tradeoffs must be considered when scaling up for real-time or resource-limited applications. Third, the PLSR model occasionally produces negative SOC values, highlighting a need for postprocessing, bounded prediction techniques or hybrid approaches. Future research should address these limitations through validation on multiple SSL datasets, hybrid modeling, and efficiency optimization.

G. Future Research

One promising research direction is optimizing NN models, which are computationally intensive because calibration requires calculating distances between all pairs of calibration samples (nc samples) for each number of components (q). Consequently, calibration complexity grows with ($nc * nc * q$) and once the optimal number of components is selected, prediction for the validation set (nv samples) involves ($nc * nv$) operations. To optimize this process, two approaches can be explored. First,

clustering the calibration samples into a smaller representative subset (e.g., reducing nc to $\sqrt{\text{nc}}$ samples) can lower complexity to approximately $\sqrt{\text{nc}} * \sqrt{\text{nc}} * q$ (i.e., nc $* q$), and validation 2 $* \sqrt{\text{nc}} *$ nv, with minimal loss in model performance. Second, create $\sqrt{\text{nc}}$ subsets of the calibration data by stratified subsetting and calibrate each separately. While it may appear more complex $\sqrt{\text{nc}} * \sqrt{\text{nc}} * \sqrt{\text{nc}} * q$ (i.e., $(\sqrt{\text{nc}})^3 * q$) for calibration and nc $*$ nv for validation, it can maintain representativeness and improve computational efficiency.

Another promising step would be to combine models to exploit their unique strengths. Evaluating model performance in a piecewise manner by stratifying the data could reveal model strengths and weaknesses under different conditions. Stratification basis can be SOC ranges (low, medium, high, very high), soil texture categories, or soil classes. This targeted evaluation may lead to more accurate predictions in specific subsets of the data. Additionally, systematically combining NN and PLSR models may balance local adaptability with broader generalization. For instance, a hybrid approach, termed NN-PLSR could first use NN model to select a locally appropriate calibration subset and then apply PLSR to derive robust, generalized regression coefficients. Such a combination could integrate the strengths of both methods, offering improved accuracy and reliability. Finally, the approaches tested in this study should be evaluated on other critical soil properties.

V. Conclusion

This study evaluated the performance of NN and PLSR models for predicting SOC across a large, globally representative, and OSSL. The results show that NN-based models, particularly *o_plsd*, can achieve superior accuracy compared to PLSR under diverse conditions. However, while *o_plsd* excels with large calibration data and effectively leverages local spectral similarities, it is more sensitive to reductions in calibration set size. By contrast, PLSR maintains more stable performance as calibration data decreases, though it is relatively more affected by the choice of sampling method. A key finding of this study is that the selection of performance metrics significantly affects model evaluation. Traditional measures like R^2, RMSE, RPD alone can be misleading, especially in large datasets with wide range of values. In this context, detailed error analysis and MAE proved more informative and robust, better reflecting true model performance. Additionally, error correlation analysis revealed structural independence, showing that certain pairs of NN models yielded redundant information; while *o_plsd* and PLSR remained distinct indicating that each model adds unique value to the prediction framework. This structural independence is crucial for informed model selection and potential development of model ensembles. Future research could explore hybrid approaches that combine the local adaptability of NN models with PLSR's broader generalization capability, potentially striking a better balance between accuracy and complexity. Furthermore, reducing computational complexity through methods like clustering calibration sets or stratified subsetting would make these models more practical for large-scale soil analysis. Ultimately, this research contributes to advancing SOC prediction methodologies, supporting more accurate and context-specific assessments critical for sustainable land management and climate change mitigation efforts.

Appendix

A. Additional Details of SOC Estimation Methods, Mathematical Approaches in Spectroscopy, the Open Soil Spectral Library (OSSL) and Stratified Sorted Sampling

This Appendix provides additional details on the variety of methods available for SOC estimation, an overview of mathematical approaches applied in soil spectroscopy, further information about the OSSL, and a description of stratified sorted sampling used in this study.

1) SOC Estimation Methods: These methods summarize various other approaches discussed in the main text. In addition to wet-chemistry techniques, other approaches for SOC estimation include visible and near-infrared spectroscopy [10], mid-infrared spectroscopy [91], covariate modelling (e.g., SCORPAN: soil properties, organisms, parent material, relief, and climate; [92]) and ecosystem simulation models (e.g., CENTURY; LPJ: Lund-Potsdam-Jena Dynamic Global Vegetation Model; CLM: Community Land Model; APEX: Agricultural Policy/Environmental eXtender; [93], [94], [95], [96]). In spectroscopy, analyzing the large datasets contained in soil spectral libraries presents significant challenges, requiring advanced methods such as dimensionality reduction, pedo-transfer functions, spectra-transfer functions, covariate analysis, transfer learning, sample subsetting, etc.

2) Open Soil Spectral Library (OSSL): The OSSL, developed by Soil Spectroscopy for Global Good, aims to provide free, transparent, and open-source research, as previously suggested by several authors [23], [24], [25], [26], [27], [97]. The spatial locations of the samples can be found in the Supplementary Material (see Fig. S2) and at the following GitHub and OSSL manual websites.[3] These diverse sources and collaborators emphasize the global and comprehensive nature of the OSSL dataset.

The OSSL includes data on 44 soil properties, sourced from ten different SSLs, including: USDA-NRCS Kellogg Soil Survey Laboratory (KSSL), ICRAF-ISRIC SSL, Africa Soil Information Service (AfSIS), LUCAS SSL, Central African Soil Spectral Library, Schiedung SSL, Garrett SSL, and Serbian SSL. These sources are cited in various studies ([23], [69], [70], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108]). Contributing organizations include the USDA NRCS National Soil Survey Center – Kellogg Soil Survey Laboratory, ICRAF-World Agroforestry, ISRIC-World Soil Information, AfSIS, European Soil Data Centre, ETH Zurich, University of Zurich, Scion Research, and University of Novi Sad, Serbia (more details can be found in the OSSL manual).

The VNIR spectra in the OSSL primarily come from the ICRAF-ISRIC SSL (4073 samples), KSSL SSL (19 807 samples), and LUCAS SSL (40 175 samples). The primary variable of interest in this study was SOC, labeled as "oc_usda.c729_w.pct," and described as "Organic Carbon, Total

[3][Online]. Available: https://github.com/soilspectroscopy; https://soilspectroscopy.github.io/ossl-manual/

TABLE III
MATHEMATICAL METHODS USED FOR SOC PREDICTION IN SOIL SPECTROSCOPY

Method	Description	Reference
MLR	Multiple Linear Regression	[109]
SMLR	Stepwise Multiple Linear Regression	[110], [111]
RT	Regression Trees	[53]
Regression-Rules	Set of linear model rules	[112]
BRT	Boosted Regression Trees	[71]
CT	Committee Trees	[53]
CART	Classification and Regression Trees	[80], [113]
MARS	Multivariate Adaptive Regression Splines	[80], [114]
PCR	Principal Component Regression	[59]
PLSR	Partial Least Squares Regression	[115]
Bagging-PLSR	Bagging Partial Least Squares Regression	[10], [76]
PLSR-ANN	Partial Least Squares Regression-Artificial Neural Networks	[116]
MWPLSR	Moving Window Partial Least Squares	[117]
Cubist	Cubist Algorithm	[37], [112], [118]
LASSO	Least Absolute Shrinkage and Selection Operator Regression	[119]
MRCE	Multivariate Regression with Covariance Estimation	[119]
CCR	Correlated Component Regression	[120]
RF	Random Forest Regression	[121], [122]
SVM/SVR and its variants	Support Vector Machines/Regression; PLS-SVR; Lap-SVR	[21], [123], [124]
DWT	Discrete Wavelet Transformation	[10]
GPR	Gaussian Process Regression	[29]
KNN	K Nearest Neighbours	[125]
NN	Nearest Neighbours	[30]
BMA	Bayesian Model Averaging	[126]
PARACUDA-2	PARACUDA-2 Data mining engine	[127]
LWR	Locally/Geographically Weighted Regression	[128]
ANN; ANN-BPNN; ANN-RBFN	Artificial Neural Networks; Artificial Neural Networks- multi-layer perceptron (MLP) feed-forward networks with a back-propagation learning algorithm.; Artificial Neural Networks - Radial basis function networks (RBFN) with regularized forward selection	[129], [130], [131]
CNN	Convolution Neural Network	[132]

TABLE IV
STRATIFIED SORTED SAMPLE SPLITS ILLUSTRATING CALIBRATION (ORANGE COLOR) AND VALIDATION (GREEN COLOR) SETS FOR VARIOUS CALIBRATION:VALIDATION RATIOS

Split ratio	Split ratio (%)	Split Nos	Calibration:Validation splits (Samples are sorted with increasing SOC content)																							
8:3	72.7:27.3	46053:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…
7:3	70.0:30.0	40297:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…
6:3	66.6:33.4	34541:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…
5:3	62.5:37.5	28785:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…
4:3	57.1:42.9	23028:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…
3:3	50.0:50.0	17271:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…
2:3	40.0:60.0	11514:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…
1:3	25.0:75.0	5757:17268	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	…

C without CaCO3, S prep." The total dataset, steps of refining and creating analysis ready dataset for this study is defined in "Section II-A Datasets."

3) Stratified Sorted Sampling: Stratified sorted sampling was employed to ensure similar SOC distributions in both calibration and validation sets. Table IV illustrates the sample splits for various calibration:validation ratios. The number in each row represents the sample number after the samples are sorted by SOC values. This approach ensures the same validation set across different ratio splits.

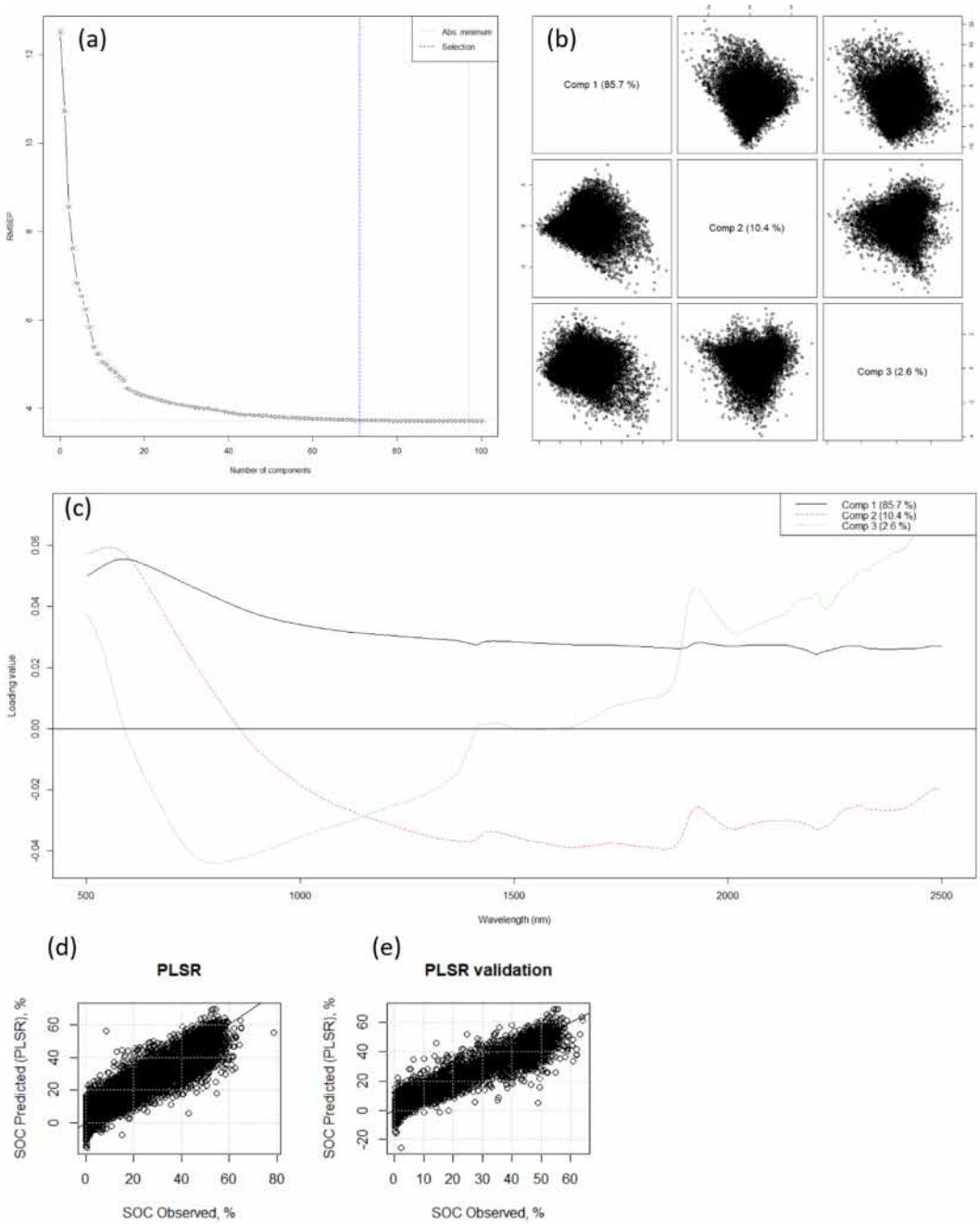


Fig. 4. PLSR model calibration and validation. (a) Selecting the number of PLS components using the one-sigma criterion. (b) Variance explained by the first three PLS components. (c) Loading values for the first three PLS components. (d) Predicted versus observed SOC for the calibration set. (e) Predicted versus observed SOC for the validation set. The plots for all calibration:validation splits are provided in supplementary information (Fig. S3-S10).

B. PLSR Calibration-Validation

The PLSR model calibration process involved selecting the optimal number of components using the one-sigma criterion to avoid an increase in RMSE (see Fig. 4). The variance explained by each PLS component and its corresponding loading factor are presented in Fig. 4(b) and (c), respectively. Predictions from the calibrated PLSR model were compared with observed SOC values, and scatter plots were generated for both the calibration and validation sets [see Fig. 4(d) and (e)]. Notably, the PLSR model sometimes produces negative SOC predictions. Such values may require further processing or cautious interpretation, particularly at very low or very high SOC levels [see Fig. 3(b) and (e); Fig. 5].

Determining the number of PLS components via the one-sigma RMSE minimization approach typically produces an elbow pattern [see Fig. 4(a)]. As the initial PLS components capture most of the spectral variance, thus rapidly reducing RMSE. Over time, as additional components explain diminishing amounts of variance (i.e., more noise than signal), reductions in RMSE taper off. In essence, the initial components represent the primary signal in the spectra, while subsequent ones capture

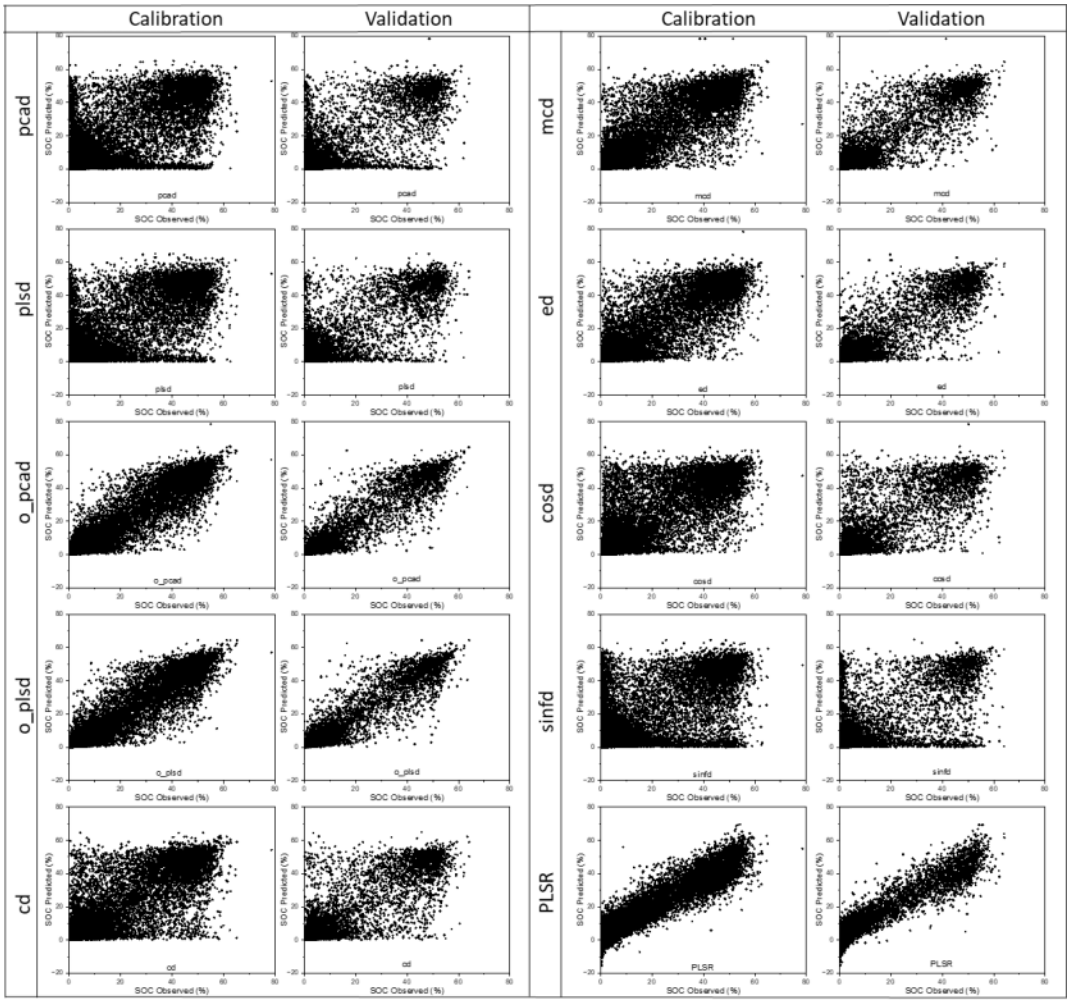


Fig. 5. Predicted versus observed SOC scatterplots for all ten models using a stratified sorted 8:3 calibration:validation split, shown for both calibration and validation sets.

progressively more noise. The loading value variations of the first three PLS components that explain 85.7%, 10.4%, and 2.6%, respectively, is shown. The variation of first component shows that all the wavelength have a positive loading value while the next two have both positive and negative loading values. The typical absorption features [see Fig. 2(a) and (b)] appear more prominently in second and third component. The first PLS component and the variation of spectra stratified with SOC value [see Figs. 2(b) and 4(c)] shows that the SOC content does not affect any particular wavelength but rather all the intensity of all the wavelengths with some more prominent than others. This could also be one of the reasons for need of many number of PLS components for prediction of SOC. The number of components needed by *o_plsd* model is 24 while for PLSR is 73 (see Table II).

C. Predictions From All Ten Models

Fig. 5 presents the scatter plots for the predicted vs. observed SOC for all methods. Models like *pcad*, *plsd*, *sinfd*, *cosd*, and *cd* show points scattered widely around the axes in both cal and val sets. In contrast, models such as *ed* and *mcd* yield predictions more closely clustered around the 1:1 line, and the best clustering is observed for *o_pcad*, *o_plsd*, and PLSR. Although PLSR's scatter plot is tightly clustered, factors like the range of predicted values and point density must also be considered. These factors may not be fully captured by simple scatter plots. Density-based scatter plots and overlaid scatter plots [see Fig. 3(b), (d), and (e)] provide clearer insights, aligning well with performance metrics like MAE (see Table II).

Additional insights are provided in Fig. 6, which shows scatter plots of model error comparisons, error distributions, and Pearson's correlation coefficients. In Fig. 6(a), the lower triangular matrix displays error scatter plots, the diagonal elements show error distributions, and the upper triangular matrix reports Pearson's correlation coefficients. These visualizations confirm the high error correlation between (*pcad*, *plsd*) and (*o_pcad*, *o_plsd*), reaffirming their structural similarity. In contrast, other model pairs remain independent [see Figs. 2(e) and 3(a)]. Thus, among the top three models identified by performance measures (*o_plsd*, *o_pcad*, PLSR), only *o_plsd* and PLSR are structurally independent.

In the NN approach, predictions for the validation set are inherently constrained by the calibration data. Since each predicted value is derived from the most similar samples in the calibration set, the predicted SOC values cannot exceed the minimum or maximum SOC values observed in that set. Consequently, the

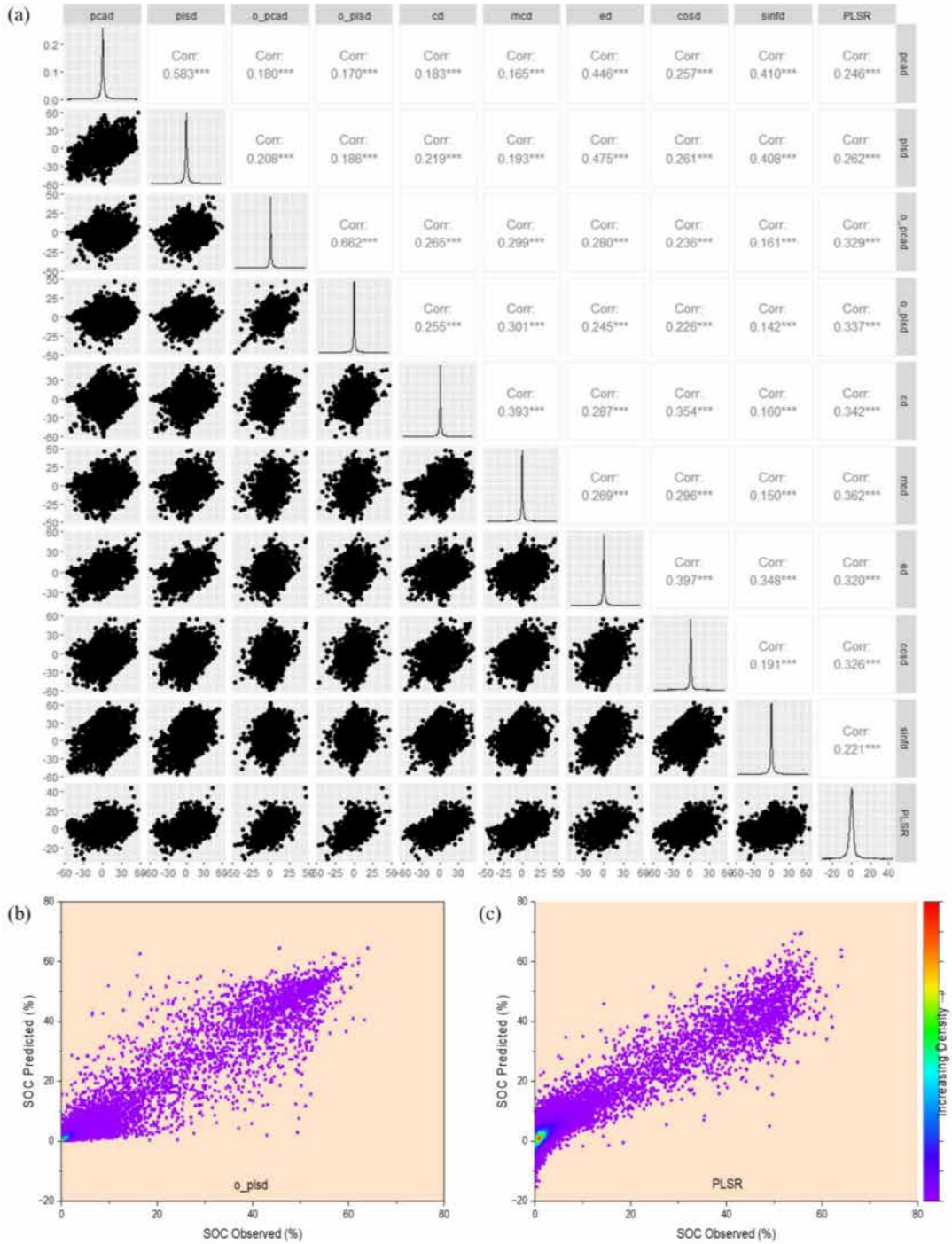


Fig. 6. Error analysis of SOC predictions including: (a) Matrix of model error comparisons, error distributions, and Pearson's correlation coefficients (lower triangular matrix displays error scatter plots, the diagonal elements show error distributions, and the upper triangular matrix reports Pearson's correlation coefficients where $***$ indicates correlation values are statistically significant at $p < 0.001$); Density-based scatter plots of predicted versus observed SOC for (b) *o_plsd* and (c) PLSR.

range of predicted SOC values in the validation set matches that of the calibration set. This inherent boundary helps prevent extreme or physically implausible SOC estimates. In contrast, PLSR does not rely on a NN framework and instead uses linear combinations of latent factors extracted from the spectra. As a result, PLSR predictions are not strictly confined to the SOC range of the calibration data and can extrapolate beyond observed values. While this flexibility may be beneficial for certain datasets, it also introduces the risk of producing unrealistic estimates. In this study, for many, PLSR predictions became negative, sometimes reaching as low as –20% SOC. Such values have no meaningful physical interpretation, emphasizing the need to apply caution when using PLSR predictions, especially at the low or high extremes of the SOC range. Figs. 3(b), 3(e),

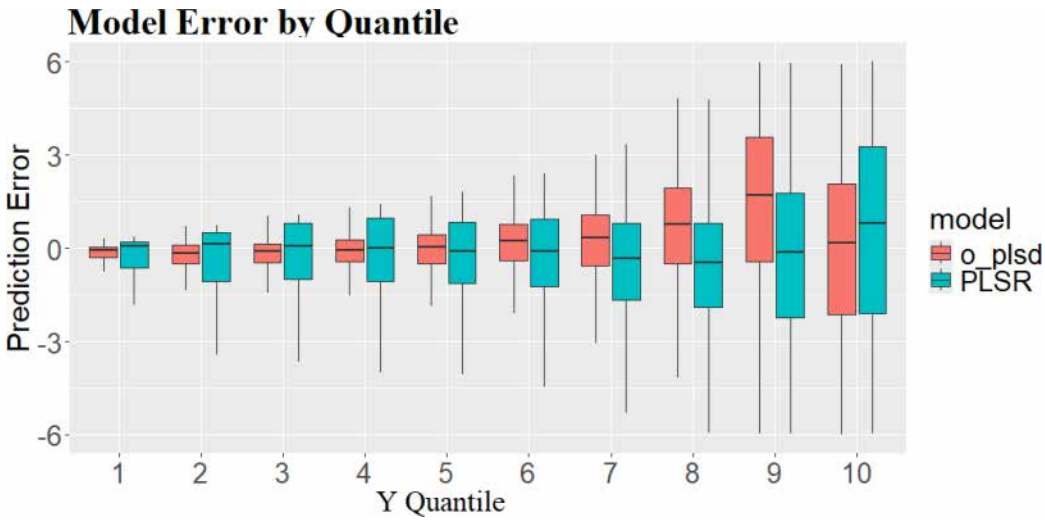


Fig. 7. Distribution of prediction errors for *o_plsd* and PLSR models across ten quantiles of SOC. Errors are computed under the 8:3 calibration:validation split scenario, with each boxplot representing the error distribution within a given SOC quantile.

TABLE V
FREQUENCY OF SAMPLES FOR WHICH *O_PLSD* AND PLSR PRODUCED THE LOWEST ABSOLUTE ERROR UNDER THE 8:3 SPLIT SCENARIO

Model	Frequency	% Frequency
o_plsd	11543	66.85
PLSR	5725	33.15
Total	**17268**	**100**

and 6 illustrate these issues, showing that while PLSR may occasionally produce extreme predictions, these outliers must be carefully scrutinized. On the other hand, the *o_plsd* model consistently yielded errors more densely concentrated near zero. Density-based scatter plots [see Fig. 6(b) and (c)] further support this finding, showing that the highest density of predicted vs. observed SOC points lies close to the origin. This pattern aligns well with the median observed SOC value of 1.91%, reflecting the model's capacity to accurately capture the central tendencies of the soil samples. In practical terms, these results suggest that *o_plsd* not only maintains a realistic range of predictions but also achieves higher overall accuracy, making it a more dependable choice for SOC estimation under a wide range of conditions.

D. Detailed Error Analysis of o_plsd and PLSR Models

Using the 8:3 stratified sorted split, we conducted an additional analysis of the validation set errors to determine, on a sample-by-sample basis, which model produced the lowest absolute error for each observation. By counting how many times each model achieved this minimum error, we can infer their relative strengths in predicting SOC values. As shown in Table V, out of a total of 17 268 validation samples, the *o_plsd* model yielded the lowest absolute error for 11 543 samples (approximately 67%), while PLSR performed best for the remaining 5725 samples (about 33%). These results indicate that *o_plsd* provides more accurate predictions for the majority of the validation samples compared to PLSR, reinforcing the conclusion that *o_plsd* is generally the stronger performer under these conditions. However, PLSR still excels for a substantial subset (one-third) of the data, highlighting that no single model is universally superior. This complementary insight, beyond aggregate performance metrics (like R^2, RMSE, RPD) can help guide model selection for specific applications. This sample-level analysis emphasizes the value of examining model performance in a more granular manner, ensuring that model selection is informed by the data characteristics and error distributions.

Furthermore, the boxplot (see Fig. 7) illustrates the distribution of prediction errors for both the models: *o_plsd* and PLSR, across ten quantiles of SOC. The *x*-axis represents the quantiles, with each bin containing 10% of the samples sorted by increasing values of SOC, while *y*-axis denotes the prediction error. Boxplots display the median, interquartile range, and overall spread of errors for each model within each quantile, with outliers excluded for clarity. In the lower quantiles, both models exhibit low error magnitudes and are centered close to zero, indicating good predictive performance. The higher SOC ranges exhibit higher error variability. The o_plsd model tends to overestimate in higher quantiles, while PLSR shows a tendency toward underestimation. This divergence in model behavior suggests that both models perform well for low to moderate SOC values but struggle to generalize accurately in higher ranges, with differing bias characteristics. Across the quantiles, *o_plsd* model consistently demonstrates lower error spread and median values closer to zero, particularly in the higher quantiles, where both models show greater variability. In contrast, the PLSR model tends to underestimate in these higher quantiles and exhibits wider error distributions.

E. Additional Details on Model Evaluation

Fig. 8 shows the methodological framework for Soil spectral library preprocessing; and model evaluation strategy for variation in input size and sampling variations. This figure is complementary to Fig. 1 with specific focus on the evaluating

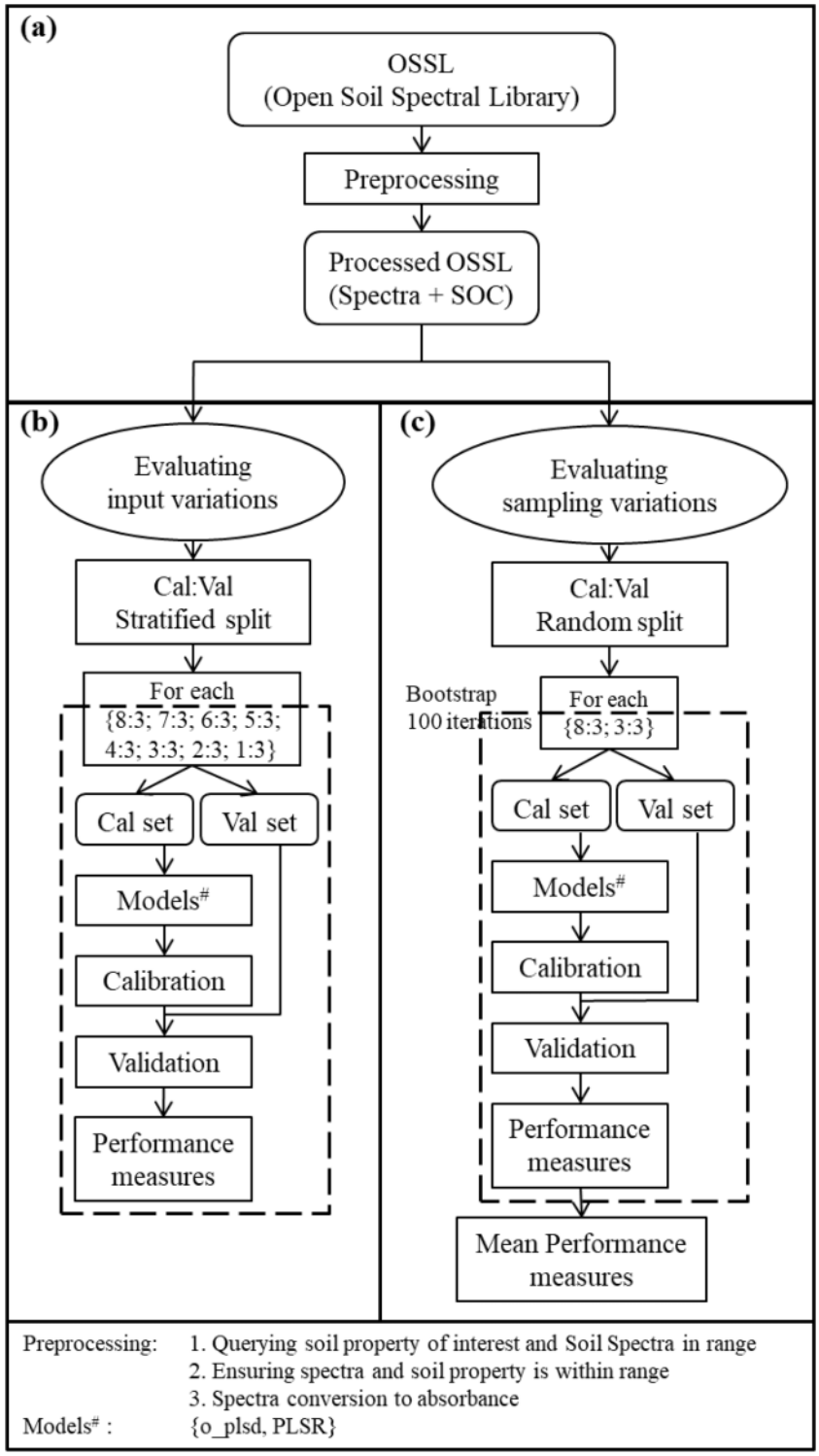


Fig. 8. Methodological framework components. (a) Soil spectral library preprocessing and generalized methodology for model evaluation including (b) evaluation of variation in input (calibration) size; (c) evaluation of variation in sampling methods.

the effects of calibration set size and sampling method on the selected models. The variation of input data should intuitively affect both *o_plsd* and PLSR models. As the calibration data size decreases, error measures are expected to increase. Similar pattern is observed in both models when data split varies from 8:3 to 1:3 (see Fig. 9). Specifically, the MAE increase from 1.79 to 2.23 in *o_plsd* and 2.36 to 2.56% in PLSR. Though the *o_plsd* model is superior to PLSR, the increase in MAE for *o_plsd* is twice that of PLSR. This implies that the *o_plsd* method is more sensitive to variation in calibration dataset size while the PLSR method is less sensitive and more generalizable. Similar conclusion could also be seen from the variation of scatter

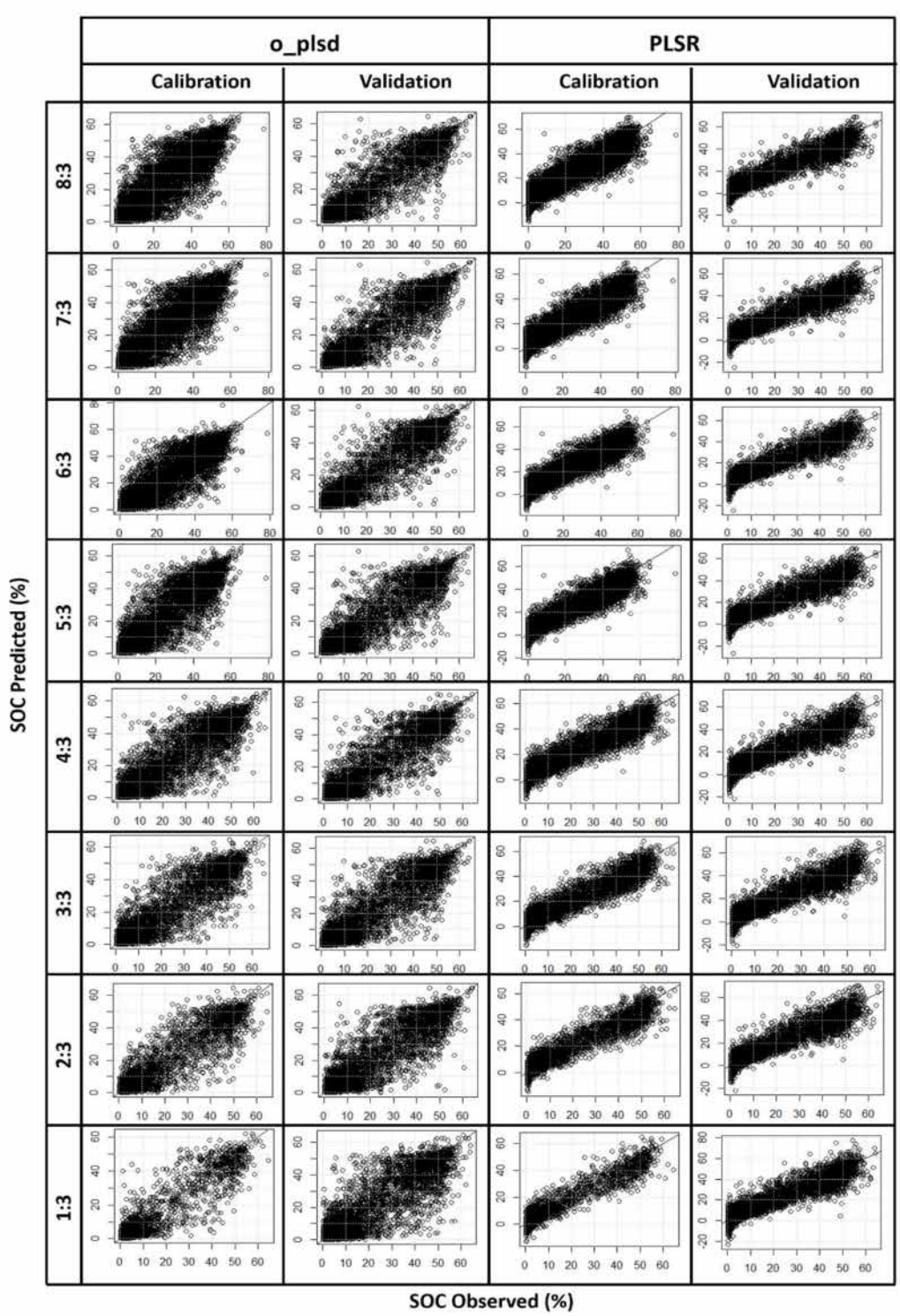


Fig. 9. Predicted versus observed SOC scatterplots for *o_plsd* and PLSR models across varying calibration:validation splits (8:3 to 1:3) in both calibration and validation sets.

plots in Fig. 9. The variation in input data size is evident from density of points in calibration scatter plots while the variation in validation plots is less evident visually. The density of points at 8:3 split is highest on 1:1 line and is reduced slightly at 1:3 split (see Fig. 9). In NN method, the range of prediction values is strictly limited by the cal set while in PLSR this is not the case. Also the range of prediction values in val set is same as that of the cal set in NN. By contrast, PLSR predictions can span a much broader range and may even produce negative SOC values, sometimes reaching as low as –20%. Such extreme values from PLSR predictions should be used cautiously, particularly at low or very high SOC concentrations [see Figs. 3(b) and (e), 9].

ACKNOWLEDGMENT

Chirag Rajendra Ternikar acknowledges discussions with Dr. Rajsekhar Kandala who introduced the concept of structural independence analysis. The authors are grateful to the OSSL team for taking the efforts over the decades to compile, quality control and open source the dataset adhering to a reproducible, clear and transparent science ethics. The authors also thank the team led by Prof. L. Ramirez-Lopez for creating easy to use "resemble" R packages along with the documentations, demonstrations and educational outreach activity.

CODE REPOSITORIES

The codes utilized in this paper along with the analysis ready data can be found on GitHub (https://github.com/ternikarcr/SOC_NN.git) and on Zenodo ([66]; https://doi.org/10.5281/zenodo.14499936).

The ways to use the packages could be found at:

https://github.com/l-ramirez-lopez/resemble?tab=readme-ov-file

https://cran.r-project.org/web/packages/resemble/vignettes/resemble.html#ref-ramirez2013spectrum

REFERENCES

[1] N. C. Brady and R. R. Weil, *The Nature and Properties of Soils*, 15th ed. Columbus, OH, USA: Pearson, 2016.

[2] R. Amundson, A. A. Berhe, J. W. Hopmans, C. Olson, A. E. Sztein, and D. L. Sparks, "Soil and human security in the 21st century," *Science*, vol. 348, no. 6235, May 2015, Art. no. 1261071, doi: 10.1126/science.1261071.

[3] H. H. Janzen, "The soil carbon dilemma: Shall we hoard it or use it?," *Soil Biol. Biochem.*, vol. 38, no. 3, pp. 419–424, Mar. 2006, doi: 10.1016/j.soilbio.2005.10.008.

[4] K. Adhikari and A. E. Hartemink, "Linking soils to ecosystem services — A global review," *Geoderma*, vol. 262, pp. 101–111, Jan. 2016, doi: 10.1016/j.geoderma.2015.08.009.

[5] P. Smith et al., "How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal," *Glob. Change Biol.*, vol. 26, no. 1, pp. 219–241, Jan. 2020, doi: 10.1111/gcb.14815.

[6] R. Lal, "Soil carbon sequestration impacts on global climate change and food security," *Science*, vol. 304, no. 5677, pp. 1623–1627, Jun. 2004, doi: 10.1126/science.1097396.

[7] A. Walkley and I. A. Black, "An examination of the degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method," *Soil Sci.*, vol. 37, no. 1, pp. 29–38, Jan. 1934, doi: 10.1097/00010694-193401000-00003.

[8] D. W. Nelson and E. L. Sommers, "Total carbon, organic carbon, and organic matter," in *Methods of Soil Analysis: Part 2 Chemical and Microbiological Properties*, vol. 9, 1982, pp. 539–579, doi: 10.2134/agronmonogr9.2.2ed.c29.

[9] M. R. Carter and E. G. Gregorich, Eds., *Soil Sampling and Methods of Analysis* . Boca Raton, FL, USA: CRC Press, 2007, doi: 10.1201/9781420005271.

[10] R. A. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no 1/2, pp. 46–54, Aug. 2010, doi: 10.1016/j.geoderma.2009.12.025.

[11] M. Nocita et al., "Soil spectroscopy: An alternative to wet chemistry for soil monitoring," in *Advances in Agronomy*, vol. 132. Amsterdam, The Netherlands: Elsevier, 2015, pp. 139–159, doi: 10.1016/bs.agron.2015.02.002.

[12] T. Angelopoulou, A. Balafoutis, G. Zalidis, and D. Bochtis, "From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—a review," *Sustainability*, vol. 12, no. 2, Jan. 2020, Art. no. 443, doi: 10.3390/su12020443.

[13] A. Ahmadi, M. Emami, A. Daccache, and L. He, "Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: A systematic review and meta-analysis," *Agronomy*, vol. 11, no. 3, Feb. 2021, Art. no. 433, doi: 10.3390/agronomy11030433.

[14] R. A. Viscarra Rossel et al., "Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century," *Eur. J. Soil Sci.*, vol. 73, no. 4, Jul. 2022, Art. no. e13271, doi: 10.1111/ejss.13271.

[15] A. V. Chinilin, G. V. Vindeker, and I. Yu. Savin, "Vis-NIR spectroscopy for soil organic carbon assessment: A meta-analysis," *Eurasian Soil Sci.*, vol. 56, no. 11, pp. 1605–1617, Nov. 2023, doi: 10.1134/S1064229323601841.

[16] A. Morón and D. Cozzolino, "Application of near infrared reflectance spectroscopy for the analysis of organic C, total N and pH in soils of Uruguay," *J. Near Infrared Spectrosc.*, vol. 10, no. 3, pp. 215–221, Jun. 2002, doi: 10.1255/jnirs.338.

[17] T. Udelhoven, C. Emmerling, and T. Jarmer, "Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study," *Plant Soil*, vol. 251, no. 2, pp. 319–329, 2003, doi: 10.1023/A:1023008322682.

[18] C. Nduwamungu, N. Ziadi, G. F. Tremblay, and L.-É. Parent, "Near-infrared reflectance spectroscopy prediction of soil properties: Effects of sample cups and preparation," *Soil Sci. Soc. Amer J.*, vol. 73, no. 6, pp. 1896–1903, Nov. 2009, doi: 10.2136/sssaj2008.0213.

[19] M. C. Sarathjith, B. S. Das, S. P. Wani, and K. L. Sahrawat, "Dependency measures for assessing the covariation of spectrally active and inactive soil properties in diffuse reflectance spectroscopy," *Soil Sci. Soc. Amer. J.*, vol. 78, no. 5, pp. 1522–1530, Sep. 2014, doi: 10.2136/sssaj2014.04.0173.

[20] A. C. Dotto et al., "Mapeamento digital de atributos: Granulometria e matéria orgânica do solo utilizando espectroscopia de reflectância difusa," *Revista Brasileira Ciência do Solo*, vol. 38, no. 6, pp. 1663–1671, Dec. 2014, doi: 10.1590/S0100-06832014000600001.

[21] Q. Jiang, Q. Li, X. Wang, Y. Wu, X. Yang, and F. Liu, "Estimation of soil organic carbon and total nitrogen in different soil layers using VNIR spectroscopy: Effects of spiking on model applicability," *Geoderma*, vol. 293, pp. 54–63, May 2017, doi: 10.1016/j.geoderma.2017.01.030.

[22] F. Lucà, M. Conforti, A. Castrignanò, G. Matteucci, and G. Buttafuoco, "Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy," *Geoderma*, vol. 288, pp. 175–183, Feb. 2017, doi: 10.1016/j.geoderma.2016.11.015.

[23] D. Garrity and P. Bindraban, "A globally distributed soil spectral library visible near infrared diffuse reflectance spectra," in *ICRAF*, (World agroforestry centre)/ISRIC (World Soil Information) Spectral Library: Nairobi, Kenya, 2004.

[24] R. A. Viscarra Rossel et al., "A global spectral library to characterize the world's soil," *Earth-Sci. Rev.*, vol. 155, pp. 198–230, Apr. 2016, doi: 10.1016/j.earscirev.2016.01.012.

[25] K. D. Shepherd, R. Ferguson, D. Hoover, F. Van Egmond, J. Sanderman, and Y. Ge, "A global soil spectral calibration library and estimation service," *Soil Secur.*, vol. 7, Jun. 2022, Art. no. 100061, doi: 10.1016/j.soisec.2022.100061.

[26] J. L. Safanelli et al., "Open soil spectral library (OSSL): Building reproducible soil calibration models through open development and community engagement," *Bioinformatics*, Dec. 2023, doi: 10.1101/2023.12.16.572011.

[27] J. L. Safanelli et al., "Open soil spectral library (OSSL): Building reproducible soil calibration models through open development and community engagement," *PLoS ONE*, vol. 20, no. 1, Jan. 2025, Art. no. e0296545, doi: 10.1371/journal.pone.0296545.

[28] F. Benedetti and FM van Egmond, "Global soil spectroscopy assessment: Spectral soil data–needs and capacities," in *Global Soil Spectroscopy Assessment: Spectral soil Data–Needs and Capacities*. Rome, Italy: FAO, 2021, doi: 10.4060/cb6265en.

[29] L. Ramirez-Lopez, T. Behrens, K. Schmidt, A. Stevens, J. A. M. Demattê, and T. Scholten, "The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets," *Geoderma*, vol. 195–196, pp. 268–279, Mar. 2013, doi: 10.1016/j.geoderma.2012.12.014.

[30] L. Ramirez-Lopez, T. Behrens, K. Schmidt, R. A. V. Rossel, J. A. M. Demattê, and T. Scholten, "Distance and similarity-search metrics for use with soil vis–NIR spectra," *Geoderma*, vol. 199, pp. 43–53, May 2013, doi: 10.1016/j.geoderma.2012.08.035.

[31] V. Bellon-Maurel, E. Fernandez-Ahumada, B. Palagos, J.-M. Roger, and A. McBratney, "Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy," *TrAC Trends Anal. Chem.*, vol. 29, no. 9, pp. 1073–1081, Oct. 2010, doi: 10.1016/j.trac.2010.05.006.

[32] W. J. M. Knoben, J. E. Freer, and R. A. Woods, "Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores," *Hydrol. Earth Syst. Sci.*, vol. 23, no. 10, pp. 4323–4331, Oct. 2019, doi: 10.5194/hess-23-4323-2019.

[33] P. H. Westfall, "Kurtosis as peakedness, 1905–2014. R.I.P.," *Amer. Statistician*, vol. 68, no. 3, pp. 191–195, Jul. 2014, doi: 10.1080/00031305.2014.917055.

[34] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, pp. 79–82, 2005, doi: 10.3354/cr030079.

[35] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.

[36] J. B. Reeves and D. B. Smith, "The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America," *Appl. Geochemistry*, vol. 24, no. 8, pp. 1472–1481, Aug. 2009, doi: 10.1016/j.apgeochem.2009.04.017.

[37] R. A. V. Rossel and R. Webster, "Predicting soil properties from the Australian soil visible–near infrared spectroscopic database," *Eur. J Soil Sci.*, vol. 63, no. 6, pp. 848–860, Dec. 2012, doi: 10.1111/j.1365-2389.2012.01495.x.

[38] B. Minasny and A. McBratney, "Why you don't need to use RPD," *Pedometron*, vol. 33, no. 600, pp. 14–15, 2013.

[39] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, 2003, doi: 10.1023/A:1022859003006.

[40] A. Gruber, W. A. Dorigo, W. Crow, and W. Wagner, "Triple collocation-based merging of satellite soil moisture retrievals," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 6780–6792, Dec. 2017, doi: 10.1109/TGRS.2017.2734070.

[41] R. Kandala, H. H. Franssen, A. Chaudhuri, and M. Sekhar, "The value of soil temperature data versus soil moisture data for state, parameter, and flux estimation in unsaturated flow model," *Vadose Zone J.*, vol. 23, no. 1, Jan. 2024, Art. no. e20298, doi: 10.1002/vzj2.20298.

[42] R. A. Viscarra Rossel, D. J. J. Walvoort, A. B. McBratney, L. J. Janik, and J. O. Skjemstad, "Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties," *Geoderma*, vol. 131, no 1/2, pp. 59–75, Mar. 2006, doi: 10.1016/j.geoderma.2005.03.007.

[43] B. Stenberg, A. Jonsson, and T. Börjesson, "Near infrared technology for soil analysis with implications for precision agriculture," in *Near Infrared Spectroscopy: Proc. 10th Int. Conf.*, Chichester, Jan. 2002, pp. 279–284.

[44] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968, doi: 10.1109/TIT.1968.1054102.

[45] P. S. Thenkabail, I. Mariotto, M. K. Gumma, E. M. Middleton, D. R. Landis, and K. F. Huemmrich, "Selection of hyperspectral narrowbands (HNBs) and composition of hyperspectral twoband vegetation indices (HVIs) for biophysical characterization and discrimination of crop types using field reflectance and hyperion/EO-1 data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 427–439, Apr. 2013, doi: 10.1109/JSTARS.2013.2252601.

[46] S. Wold, J. Trygg, A. Berglund, and H. Antti, "Some recent developments in PLS modeling," *Chemometrics Intell. Lab. Syst.*, vol. 58, no. 2, pp. 131–150, Oct. 2001, doi: 10.1016/S0169-7439(01)00156-3.

[47] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Analytica Chim. Acta*, vol. 185, pp. 1–17, 1986, doi: 10.1016/0003-2670(86)80028-9.

[48] L. Ramirez-Lopez, A. Stevens, R. Viscarra Rossel, Z. Shen, A. Wadoux, and T. Breure, "Resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics. R package vignette," R package version 2.2.3, 2024.

[49] S. R. S. Dangal, J. Sanderman, S. Wills, and L. Ramirez-Lopez, "Accurate and precise prediction of soil properties from a large mid-infrared spectral library," *Soil Syst.*, vol. 3, no. 1, Jan. 2019, Art. no. 11, doi: 10.3390/soilsystems3010011.

[50] C. M. Clingensmith and S. Grunwald, "Predicting soil properties and interpreting Vis-NIR models from across continental United States," *Sensors*, vol. 22, no. 9, Apr. 2022, Art. no. 3187, doi: 10.3390/s22093187.

[51] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Stat. Learn.: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2013.

[52] B. H. Mevik and R. Wehrens, "The pls package: Principal component and partial least squares regression in R," *J. Stat. Soft.*, vol. 18, no. 2, pp. 1–23, Jan. 2007, doi: 10.18637/jss.v018.i02.

[53] G. M. Vasques, S. Grunwald, and J. O. Sickman, "Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra," *Geoderma*, vol. 146, no 1/2, pp. 14–25, Jul. 2008, doi: 10.1016/j.geoderma.2008.04.007.

[54] A. Gupta, H. B. Vasava, B. S. Das, and A. K. Choubey, "Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region," *Geoderma*, vol. 325, pp. 59–71, Sep. 2018, doi: 10.1016/j.geoderma.2018.03.025.

[55] H. B. Vasava, A. Gupta, R. Arora, and Bhabani. S. Das, "Assessment of soil texture from spectral reflectance data of bulk soil samples and their dry-sieved aggregate size fractions," *Geoderma*, vol. 337, pp. 914–926, Mar. 2019, doi: 10.1016/j.geoderma.2018.11.004.

[56] W. Osterholz, K. King, M. Williams, B. Hanrahan, and E. Duncan, "Stratified soil sampling improves predictions of P concentration in surface runoff and tile discharge," *Soil Syst.*, vol. 4, no. 4, Nov. 2020, Art. no. 67, doi: 10.3390/soilsystems4040067.

[57] E. B. George, C. Gomez, D. Nagesh Kumar, S. Dharumarajan, and M. Lalitha, "Impact of bare soil pixels identification on clay content mapping using airborne hyperspectral AVIRIS-NG data: Spectral indices *versus* spectral unmixing," *Geocarto Int.*, vol. 37, no. 27, pp. 15912–15934, Dec. 2022, doi: 10.1080/10106049.2022.2102241.

[58] G. M. Vasques, S. Grunwald, and W. G. Harris, "Spectroscopic models of soil organic carbon in Florida, USA," *J. Environ. Qual.*, vol. 39, no. 3, pp. 923–934, May 2010, doi: 10.2134/jeq2009.0314.

[59] C.-W. Chang, D. A. Laird, M. J. Mausbach, and C. R. Hurburgh, "Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties," *Soil Sci. Soc. Amer. J.*, vol. 65, no. 2, pp. 480–490, Mar. 2001, doi: 10.2136/sssaj2001.652480x.

[60] B. Stenberg, R. A. Viscarra Rossel, A. M. Mouazen, and J. Wetterlind, "Visible and near infrared spectroscopy in soil science," in *Advances in Agronomy*, vol. 107. Amsterdam, The Netherlands: Elsevier, 2010, pp. 163–215, doi: 10.1016/S0065-2113(10)07005-7.

[61] M. Conforti, G. Buttafuoco, A. P. Leone, P. P. Aucelli, G. Robustelli, and F. Scarciglia, "Studying the relationship between water-induced soil erosion and soil organic matter using Vis–NIR spectroscopy and geomorphological analysis: A case study in southern Italy," *CATENA*, vol. 110, pp. 44–58, 2013.

[62] J. M. Moura-Bueno, R. S. D. Dalmolin, A. Ten Caten, A. C. Dotto, and J. A. M. Demattê, "Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions," *Geoderma*, vol. 337, pp. 565–581, Mar. 2019, doi: 10.1016/j.geoderma.2018.10.015.

[63] J. K. Carvalho et al., "Combining different pre-processing and multivariate methods for prediction of soil organic matter by near infrared spectroscopy (NIRS) in Southern Brazil," *Geoderma Regional*, vol. 29, Jun. 2022, Art. no. e00530, doi: 10.1016/j.geodrs.2022.e00530.

[64] M. Conforti and G. Buttafuoco, "Insights into the effects of study area size and soil sampling density in the prediction of soil organic carbon by Vis-NIR diffuse reflectance spectroscopy in two forest areas," *Land*, vol. 12, no. 1, Dec. 2022, Art. no. 44, doi: 10.3390/land12010044.

[65] R Core Team, "R A language and environment for statistical computing," *R Found. Stat. Comput.*, 2022.

[66] C. R. Ternikar, C. Gomez, D. Dutta, and D. N. Kumar, "Ternikarcr/SOC_NN: SOC estimation using nearest neighbour models," Zenodo, Oct. 2024, doi: 10.5281/ZENODO.14499936.

[67] A. G. Asuero, A. Sayago, and A. G. González, "The correlation coefficient: An overview," *Crit. Rev. Anal. Chem.*, vol. 36, no. 1, pp. 41–59, Jan. 2006, doi: 10.1080/10408340500526766.

[68] W. Saeys, A. M. Mouazen, and H. Ramon, "Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy," *Biosyst. Eng.*, vol. 91, no. 4, pp. 393–402, Aug. 2005, doi: 10.1016/j.biosystemseng.2005.05.001.

[69] N. K. Wijewardane, Y. Ge, S. Wills, and T. Loecke, "Prediction of soil carbon in the conterminous United States: Visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project," *Soil Sci. Soc. Amer. J.*, vol. 80, no. 4, pp. 973–982, Jul. 2016, doi: 10.2136/sssaj2016.02.0052.

[70] S. Wills, T. Loecke, C. Sequeira, G. Teachman, S. Grunwald,, and L. T. West, "Overview of the US rapid carbon assessment project: Sampling design, initial summary and uncertainty estimates," *Soil Carbon*, pp. 95–104, 2014, doi: 10.1007/978-3-319-04084-4_10.

[71] D. J. Brown, K. D. Shepherd, M. G. Walsh, M. Dewayne Mays, and T. G. Reinsch, "Global soil characterization with VNIR diffuse reflectance spectroscopy," *Geoderma*, vol. 132, no 3/4, pp. 273–290, Jun. 2006, doi: 10.1016/j.geoderma.2005.04.025.

[72] A. Stevens, M. Nocita, G. Tóth, L. Montanarella, and B. Van Wesemael, "Prediction of soil organic carbon at the European scale by visible and near InfraRed reflectance spectroscopy," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e66409, doi: 10.1371/journal.pone.0066409.

[73] T. Shi, Y. Chen, H. Liu, J. Wang, and G. Wu, "Soil organic carbon content estimation with laboratory-based visible–near-infrared reflectance spectroscopy: Feature selection," *Appl. Spectrosc.*, vol. 68, no. 8, pp. 831–837, Aug. 2014, doi: 10.1366/13-07294.

[74] F. S. Terra, J. A. M. Demattê, and R. A. Viscarra Rossel, "Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data," *Geoderma*, vol. 255–256, pp. 81–93, Oct. 2015, doi: 10.1016/j.geoderma.2015.04.017.

[75] G. Shao, L. Tang, and J. Liao, "Overselling overall map accuracy misinforms about research reliability," *Landscape Ecol.*, vol. 34, no. 11, pp. 2487–2492, Nov. 2019, doi: 10.1007/s10980-019-00916-6.

[76] R. A. V. Rossel, Y. S. Jeon, I. O. A. Odeh, and A. B. McBratney, "Using a legacy soil sample to develop a mid-IR spectral library," *Soil Res.*, vol. 46, no. 1, 2008, Art. no. 1, doi: 10.1071/SR07099.

[77] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomic.*, vol. 21, no. 1, Dec. 2020, Art. no. 6, doi: 10.1186/s12864-019-6413-7.

[78] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, Mar. 2024, Art. no. 6086, doi: 10.1038/s41598-024-56706-x.

[79] W. Ng, B. Minasny, B. Malone, and P. Filippi, "In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra," *PeerJ*, vol. 6, Oct. 2018, Art. no. e5722, doi: 10.7717/peerj.5722.

[80] K. D. Shepherd and M. G. Walsh, "Development of reflectance spectral libraries for characterization of soil properties," *Soil Sci. Soc. Amer. J.*, vol. 66, no. 3, pp. 988–998, 2002, doi: 10.2136/sssaj2002.9880.

[81] B. Kuang and A. M. Mouazen, "Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale," *Eur. J. Soil Sci.*, vol. 63, no. 3, pp. 421–429, Jun. 2012, doi: 10.1111/j.1365-2389.2012.01456.x.

[82] C. Grinand et al., "Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS)," *Eur. J. Soil Sci.*, vol. 63, no. 2, pp. 141–151, Apr. 2012, doi: 10.1111/j.1365-2389.2012.01429.x.

[83] G. Debaene, J. Niedźwiecki, A. Pecio, and A. Żurek, "Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale," *Geoderma*, vol. 214–215, pp. 114–125, Feb. 2014, doi: 10.1016/j.geoderma.2013.09.022.

[84] F. Gogé, C. Gomez, C. Jolivet, and R. Joffre, "Which strategy is best to predict soil properties of a local site from a national Vis–NIR database?," *Geoderma*, vol. 213, pp. 1–9, Jan. 2014, doi: 10.1016/j.geoderma.2013.07.016.

[85] L. Ramirez-Lopez, K. Schmidt, T. Behrens, B. Van Wesemael, J. A. M. Demattê, and T. Scholten, "Sampling optimal calibration sets in soil infrared spectroscopy," *Geoderma*, vol. 226–227, pp. 140–150, Aug. 2014, doi: 10.1016/j.geoderma.2014.02.002.

[86] M. Clairotte et al., "National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy," *Geoderma*, vol. 276, pp. 41–52, Aug. 2016, doi: 10.1016/j.geoderma.2016.04.021.

[87] C. M. Clingensmith, S. Grunwald, and S. P. Wani, "Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment," *Eur. J. Soil Sci.*, vol. 70, no. 1, pp. 107–126, Jan. 2019, doi: 10.1111/ejss.12753.

[88] M. J. Dorantes, B. A. Fuentes, and D. M. Miller, "Calibration set optimization and library transfer for soil carbon estimation using soil spectroscopy—A review," *Soil Sci. Soc. Amer. J.*, vol. 86, no. 4, pp. 879–903, Jul. 2022, doi: 10.1002/saj2.20435.

[89] D. J. Brown, R. S. Bricklemyer, and P. R. Miller, "Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana," *Geoderma*, vol. 129, no. 3/4, pp. 251–267, Dec. 2005, doi: 10.1016/j.geoderma.2005.01.001.

[90] J. M. Soriano-Disla, L. J. Janik, R. A. Viscarra Rossel, L. M. Macdonald, and M. J. McLaughlin, "The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties," *Appl. Spectrosc. Rev.*, vol. 49, no. 2, pp. 139–186, Feb. 2014, doi: 10.1080/05704928.2013.811081.

[91] G. W. McCarty, J. B. Reeves, V. B. Reeves, R. F. Follett, and J. M. Kimble, "Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement," *Soil Sci. Soc. Amer. J.*, vol. 66, no. 2, pp. 640–646, Mar. 2002, doi: 10.2136/sssaj2002.6400a.

[92] A. B. McBratney, M. L. Mendonça Santos, and B. Minasny, "On digital soil mapping," *Geoderma*, vol. 117, no 1/2, pp. 3–52, Nov. 2003, doi: 10.1016/S0016-7061(03)00223-4.

[93] W. Parton et al., "Global-scale similarities in nitrogen release patterns during long-term decomposition," *Science*, vol. 315, no. 5810, pp. 361–364, Jan. 2007, doi: 10.1126/science.1134853.

[94] S. Sitch et al., "Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model," *Glob. Change Biol.*, vol. 9, no. 2, pp. 161–185, Feb. 2003, doi: 10.1046/j.1365-2486.2003.00569.x.

[95] J. R. Williams and R. C. Izaurralde, *The APEX Model in Watershed Models*. Boca Raton, FL, USA: CRC Press, 2010.

[96] D. M. Lawrence et al., "Parameterization improvements and functional and structural advances in Version 4 of the community land model: Parameterization improvements and functional and structural advances," *J. Adv. Model. Earth Syst.*, vol. 3, no. 1, Jan. 2011, doi: 10.1029/2011MS00045.

[97] J. L. Safanelli et al., "An interlaboratory comparison of mid-infrared spectra acquisition: Instruments and procedures matter," *Geoderma*, vol. 440, Dec. 2023, Art. no. 116724, doi: 10.1016/j.geoderma.2023.116724.

[98] J. Sanderman, K. Savage, and S. R. S. Dangal, "Mid-infrared spectroscopy for prediction of soil health indicators in the United States," *Soil Sci. Soc. Amer. J.*, vol. 84, no. 1, pp. 251–261, Jan. 2020, doi: 10.1002/saj2.20009.

[99] N. K. Wijewardane, Y. Ge, S. Wills, and Z. Libohova, "Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library," *Soil Sci. Soc. Amer. J.*, vol. 82, no. 3, pp. 722–731, May 2018, doi: 10.2136/sssaj2017.10.0361.

[100] M. J. Aitkenhead and H. I. J. Black, "Exploring the impact of different input data types on soil variable estimation using the ICRAF-ISRIC global soil spectral database," *Appl. Spectrosc.*, vol. 72, no. 2, pp. 188–198, Feb. 2018, doi: 10.1177/0003702817739013.

[101] A. Jones, o Fernandez-Ugalde, and S. Scarpa, "LUCAS 2015 topsoil survey. Presentation of dataset and results," EUR 30332, 2020.

[102] T. G. Vagen et al., "Mid-infrared spectra (MIRS) from ICRAF soil and plant spectroscopy laboratory: Africa soil information service (AfSIS) phase I 2009–2013," World Agroforestry–Research Data Repository, 1, 2020.

[103] T. Hengl et al., "African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning," *Sci. Rep.*, vol. 11, no. 1, Mar. 2021, Art. no. 6130, doi: 10.1038/s41598-021-85639-y.

[104] A. Orgiazzi, C. Ballabio, P. Panagos, A. Jones, and O. Fernández-Ugalde, "LUCAS Soil, the largest expandable soil dataset for Europe: A review," *Eur. J. Soil Sci.*, vol. 69, no. 1, pp. 140–153, Jan. 2018, doi: 10.1111/ejss.12499.

[105] L. Summerauer et al., "The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis," *SOIL*, vol. 7, no. 2, pp. 693–715, Oct. 2021, doi: 10.5194/soil-7-693-2021.

[106] M. Schiedung, S.-L. Bellè, A. Malhotra, and S. Abiven, "Organic carbon stocks, quality and prediction in permafrost-affected forest soils in North Canada," *CATENA*, vol. 213, Jun. 2022, Art. no. 106194, doi: 10.1016/j.catena.2022.106194.

[107] L. G. Garrett et al., "Mid-infrared spectroscopy for planted forest soil and foliage nutrition predictions, New Zealand case study," *Trees, Forests People*, vol. 8, Jun. 2022, Art. no. 100280, doi: 10.1016/j.tfp.2022.100280.

[108] B. Jović, V. Ćirić, M. Kovačević, S. Šeremešić, and B. Kordić, "Empirical equation for preliminary assessment of soil texture," *Spectrochim. Acta Part A, Mol. Biomol. Spectrosc.*, vol. 206, pp. 506–511, Jan. 2019, doi: 10.1016/j.saa.2018.08.039.

[109] E. Ben-Dor and A. Banin, "Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties," *Soil Sci. Soc. Amer. J.*, vol. 59, no. 2, pp. 364–372, Mar. 1995, doi: 10.2136/sssaj1995.03615995005900020014x.

[110] R. C. Dalal and R. J. Henry, "Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry," *Soil Sci. Soc. Amer. J.*, vol. 50, no. 1, pp. 120–123, Jan. 1986, doi: 10.2136/sssaj1986.03615995005000010023x.

[111] D. F. Malley, P. D. Martin, L. M. McClintock, L. Yesmin, R. G. Eilers, and P. Haluschak, "Feasibility of analysing archived Canadian prairie agricultural soils by near infrared reflectance spectroscopy," in *Near Infrared Spectroscopy: Proc. 9th Int. Conf.*, Chichester, U.K., Jun., 2000, pp. 579–585.

[112] B. Minasny and A. B. McBratney, "Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy," *Chemometrics Intell. Lab. Syst.*, vol. 94, no. 1, pp. 72–79, Nov. 2008, doi: 10.1016/j.chemolab.2008.06.003.

[113] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*, 1st ed. Evanston, IL, USA: Routledge, 2017, doi: 10.1201/9781315139470.
[114] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Statist.*, vol. 19, no. 1, pp. 1–67, 1991.
[115] K. A. Sudduth and J. W. Hummel, "Soil organic matter, CEC, and moisture sensing with a portable NIR spectrophotometer," *Trans. ASAE*, vol. 36, no. 6, pp. 1571–1582, 1993, doi: 10.13031/2013.28498.
[116] A. M. Mouazen, B. Kuang, J. De Baerdemaeker, and H. Ramon, "Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy," *Geoderma*, vol. 158, no 1/2, pp. 23–31, Aug. 2010, doi: 10.1016/j.geoderma.2010.03.001.
[117] H. Chen, T. Pan, J. Chen, and Q. Lu, "Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods," *Chemometrics Intell. Lab. Syst.*, vol. 107, no. 1, pp. 139–146, May 2011, doi: 10.1016/j.chemolab.2011.02.008.
[118] J. R. Quinlan, "Learning with continuous classes," in *Proc. 5th Australian joint Conf. Artif. Intell.*, Nov. 1992, pp. 343–348.
[119] R. S. Bricklemyer, D. J. Brown, P. J. Turk, and S. Clegg, "Comparing vis–NIRS, LIBS, and combined vis–NIRS-LIBS for intact soil core soil carbon measurement," *Soil Sci. Soc. Amer. J.*, vol. 82, no. 6, pp. 1482–1496, Nov. 2018, doi: 10.2136/sssaj2017.09.0332.
[120] O. A. Rosero-Vlasova, L. Vlassova, F. Pérez-Cabello, R. Montorio, and E. Nadal-Romero, "Soil organic matter and texture estimation from visible–near infrared–shortwave infrared spectra in areas of land cover changes using correlated component regression," *Land Degradation Develop.*, vol. 30, no. 5, pp. 544–560, Mar. 2019, doi: 10.1002/ldr.3250.
[121] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[122] A. Gholizadeh et al., "Soil organic carbon estimation using VNIR–SWIR spectroscopy: The effect of multiple sensors and scanning conditions," *Soil Tillage Res.*, vol. 211, Jul. 2021, Art. no. 105017, doi: 10.1016/j.still.2021.105017.
[123] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.
[124] N. L. Tsakiridis, J. B. Theocharis, A. L. Symeonidis, and G. C. Zalidis, "Improving the predictions of soil properties from VNIR–SWIR spectra in an unlabeled region using semi-supervised and active learning," *Geoderma*, vol. 387, Apr. 2021, Art. no. 114830, doi: 10.1016/j.geoderma.2020.114830.
[125] K. M. D. Oliveira et al., "Predicting particle size and soil organic carbon of soil profiles using VIS-NIR-SWIR hyperspectral imaging and machine learning models," *Remote Sens.*, vol. 16, no. 16, Aug. 2024, Art. no. 2869, doi: 10.3390/rs16162869.
[126] A. C. Dotto, R. S. D. Dalmolin, A. Ten Caten, and S. Grunwald, "A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra," *Geoderma*, vol. 314, pp. 262–274, Mar. 2018, doi: 10.1016/j.geoderma.2017.11.006.
[127] A. Gholizadeh, D. Žižala, M. Saberioon, and L. Borůvka, "Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging," *Remote Sens. Environ.*, vol. 218, pp. 89–103, Dec. 2018, doi: 10.1016/j.rse.2018.09.015.
[128] Y. Song, Z. Shen, P. Wu, and R. A. Viscarra Rossel, "Wavelet geographically weighted regression for spectroscopic modelling of soil properties," *Sci. Rep.*, vol. 11, no. 1, Sep. 2021, Art. no. 17503, doi: 10.1038/s41598-021-96772-z.
[129] P. H. Fidêncio, R. J. Poppi, and J. C. De Andrade, "Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy," *Analytica Chim. Acta*, vol. 453, no. 1, pp. 125–134, Feb. 2002, doi: 10.1016/S0003-2670(01)01506-9.
[130] K. W. Daniel, N. K. Tripathi, and K. Honda, "Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand)," *Soil Res.*, vol. 41, no. 1, 2003, Art. no. 47, doi: 10.1071/SR02027.
[131] S. Katuwal, M. Knadel, T. Norgaard, P. Moldrup, M. H. Greve, and L. W. De Jonge, "Predicting the dry bulk density of soils across Denmark: Comparison of single-parameter, multi-parameter, and vis–NIR based models," *Geoderma*, vol. 361, Mar. 2020, Art. no. 114080, doi: 10.1016/j.geoderma.2019.114080.
[132] W. Ng et al., "Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra," *Geoderma*, vol. 352, pp. 251–267, Oct. 2019, doi: 10.1016/j.geoderma.2019.06.016.

Chirag Rajendra Ternikar received the B.Tech. degree in civil engineering from the Walchand College of Engineering, Sangli, India, in 2016, the M.Tech. degree in remote sensing & GIS from NIT Surathkal, Karnataka, India, in 2018. He is currently working toward the Ph.D. degree in water resources and environmental engineering from the Department of Civil Engineering, Indian Institute of Science, Bangalore, India.

His research interests include soil property estimation using laboratory, airborne and space-borne remote sensing measurements. Other research interests include data assimilation techniques for water budget closure, evapotranspiration estimation, above ground biomass estimation and advanced technique development for unmixing and classification of hyperspectral data.

Cécile Gomez received the Ph.D. degree in earth sciences from the Laboratoire Sciences de la Terre, Lyon, France, in 2004.

She is a Researcher with the French Institute of Research for Development, Montpellier, France. She is part of the Laboratory on Interactions between Soil, Agrosystem and Hydrosystem (LISAH) and also the Indo-French cell for Water Science, at the Indian Institute of Sciences in Bangalore (India) from 2019. Her main research interests include VNIR/SWIR spectroscopy and remote sensing data treatments (hyperspectral and multispectral) for soil mapping.

Debsunder Dutta received the B.Eng. degree from the Indian Institute of Engineering Science and Technology, Shibpur, India, in 2009, the M.Tech. degree from IIT Kanpur, Kanpur, India, in 2011, and the Ph.D. degree from the University of Illinois at Urbana–Champaign, Urbana, IL, USA, in 2016, all in civil engineering.

He is currently an Assistant Professor with the Department of Civil Engineering, Indian Institute of Science, Bangalore, India, prior to which he was a Postdoctoral Fellow with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. His research interests include the applications of high-resolution remote sensing measurements to better understand eco-hydrologic processes, with a focus on using hyperspectral and LiDAR data for the quantitative estimation of soil and vegetation attributes using data-driven approaches.

D. Nagesh Kumar (Senior Member, IEEE) received the Ph.D. degree in civil engineering from Indian Institute of Science, Bangalore, India, in 1992.

He is working as a Professor with the Department of Civil Engineering, Indian Institute of Science, Bangalore, India, since May 2002. He is on sabbatical from IISc during 2024–2025 and working as a Curtis Visiting Professor with the Lyles School of Civil and Construction Engineering, Purdue University, West Lafayette, IN, USA. His research interests include Climate Hydrology, Water Resources Systems, ANN, Evolutionary Algorithms, Fuzzy logic, MCDM and Remote Sensing & GIS applications in water resources engineering. He has coauthored two text books *Multicriterion Analysis in Engineering and Management* (PHI, New Delhi) and *Floods in a Changing Climate: Hydrologic Modeling* (Cambridge University Press, U.K.). He has supervised 10 Postdocs (1 in progress) and 22 Ph.D. (7 in progress). He has coauthored 8 books and published more than 225 papers including 136 in peer reviewed journals. He has received funding support of more than Rs. 35 crores for sponsored research.

Mr. Kumar is the Editor in Chief of *Journal of Water and Climate Change*, IWA Publishing, U.K., and an Associate Editor for *ASCE Journal of Hydrologic Engineering*. He was the recipient of IBM Faculty Award for his outstanding contributions in modelling hydrologic extremes using microwave remote sensing.