https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

Evaluating Vis–NIR spectroscopy for laboratory and in-situ prediction of forest soil organic carbon fractions

Jiaxin Li [a, b, d], Pierre Roudier [c], Wenli Fei [a], Lidu Shen [a], Yage Liu [a], Anzhi Wang [a], Sam McNally [d], Yuan Zhang [a, *], Jiabing Wu [a, *]

a CAS Key Laboratory of Forest Ecology and Silviculture, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China

b University of Chinese Academy of Sciences, Beijing, China

c Bioeconomy Science Institute, Manaaki Whenua – Landcare Research, Palmerston North, New Zealand

d Bioeconomy Science Institute, Manaaki Whenua – Landcare Research group, Lincoln, New Zealand

* *Corresponding author.* E-mail address: zhangyuan@iae.ac.cn; wujb@iae.ac.cn

Abstract

Forest soil organic carbon (SOC) stability is influenced by its relative composition of particulate organic carbon (POC) and mineral-associated organic carbon (MAOC) fractions. However, conventional SOC fractionation methods are labor-intensive and restrict large-scale monitoring of SOC dynamics. Visible–near infrared (Vis–NIR) spectroscopy offers a rapid alternative, yet its applicability for predicting SOC fractions in forest soils under field conditions remains poorly understood. This study developed an integrated framework to evaluate the feasibility of in-situ Vis–NIR spectroscopy for predicting SOC fractions by comparing four in-situ application workflows, including direct laboratory-to-field transfer, EPO-assisted transfer, direct in-situ modeling, and EPO-assisted in-situ modeling. Direct transfer of laboratory models to in-situ spectra resulted in substantial performance degradation due to moisture-driven spectral domain shifts (POC: $R^2 = 0.80$; MAOC: $R^2 = 0.59$). In contrast, direct in-situ modeling.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

The highest accuracy for POC was achieved using EPO-corrected in-situ spectra (R^2 = 0.90), whereas MAOC prediction performed best using uncorrected in-situ spectra (R^2 = 0.71). Independent cross-year validation further demonstrated that environmental variability, particularly soil moisture, constrained model robustness. The analysis of the fitted models revealed distinct spectral mechanisms controlling SOC fraction predictions, linking POC to shortwave infrared organo–clay absorption features (~2200 nm) and MAOC to visible wavelengths associated with iron oxides. These findings highlight the conditional feasibility of in-situ Vis–NIR spectroscopy for forest SOC fraction prediction and guide field-based soil carbon monitoring.

Keywords: Spectral transferability; External parameter orthogonalization; Particulate organic carbon; Mineral-associated organic carbon; In-situ spectroscopy; Model robustness

Introduction

Soil organic carbon (SOC) plays a central role in regulating the global carbon cycle, and forest soils constitute one of the largest terrestrial carbon reservoirs. However, the stability and persistence of SOC are governed not only by its total stock but also by the behavior of its constituent fractions. Increasing evidence indicates that the response of SOC to environmental change is strongly controlled by the relative proportions of particulate organic carbon (POC) and mineral-associated organic carbon (MAOC), which differ fundamentally in their formation pathways, residence time, persistence, and ecological functions (Lavallee et al., 2020; Zhou et al., 2024). POC represents a relatively labile carbon pool derived primarily

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

from partially decomposed plant residues and microbial products, whereas MAOC forms a more stable carbon pool through associations with mineral surfaces. Consequently, the distribution of these fractions provides key insights into soil carbon stability, vulnerability to loss, and long-term sequestration potential (Abramoff et al., 2022; Viscarra Rossel et al., 2024).

Despite their importance, quantifying SOC fractions remains challenging. Conventional approaches rely mainly on physical, chemical, or thermal fractionation techniques (Delahaie et al., 2024), with physical fractionation being the most widely used (Wu et al., 2025). Although conceptually robust, these methods are labor-intensive, time-consuming, and costly, limiting their applicability for large-scale monitoring or high-frequency assessments of soil carbon dynamics (De Vos et al., 2015). Developing rapid and scalable approaches to estimate SOC fractions is therefore essential to advancing soil carbon research and monitoring.

Diffuse reflectance spectroscopy in the visible–near infrared range (Vis–NIR, 350 to 2,500 nm) provides a promising alternative for rapid soil analysis. Because soil organic matter and minerals exhibit characteristic absorption features across the Vis–NIR region, this technique enables the simultaneous estimation of multiple soil properties (R. A. Viscarra Rossel et al., 2006; Rossel and Behrens, 2010). Over the past two decades, laboratory-based Vis–NIR spectroscopy has demonstrated strong predictive capability for bulk SOC and other soil properties (Jaconi et al., 2017). However, the application of Vis–NIR spectroscopy to SOC fractions remains comparatively limited. Existing studies have predominantly relied on mid-infrared spectroscopy to capture fundamental vibrational features of organic functional groups and minerals (Janik et al., 2007; Pacini et al., 2025), whereas relatively few have explored the potential of Vis–NIR spectroscopy to predict SOC fractions (Qi et al., 2024).

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

While most current spectral models for SOC fractions were developed using air-dried soils measured under controlled laboratory conditions, the recent emergence of portable Vis–NIR spectrometers offers a practical solution for acquiring soil spectra directly in the field (Conforti et al., 2018; Silvero et al., 2020). In contrast to laboratory spectroscopy, spectra recorded under field conditions are strongly influenced by environmental factors such as soil moisture, surface roughness, and sensor–soil contact conditions, which can substantially alter spectral signatures and reduce the transferability of laboratory-based models (Roudier et al., 2017; Koch et al., 2021). This challenge is even more pronounced in forest ecosystems when acquiring soil reflectance data via airborne or satellite remote sensing, as dense canopy cover and litter layers often prevent the spectral signal from reaching the soil surface. Consequently, the reliability of in-situ Vis–NIR spectroscopy for predicting SOC fractions in forest soils remains poorly understood.

Several methodological strategies have been proposed to address the discrepancy between laboratory and in-situ spectra. One approach involves directly transferring laboratory-calibrated models to in-situ spectra, while another involves calibrating models with in-situ measurements. In addition, spectral correction techniques such as external parameter orthogonalization (EPO) have been proposed to mitigate environmental interference by removing spectral variance attributable to external factors, such as soil moisture (Minasny et al., 2011; Roudier et al., 2017). However, the relative effectiveness of these strategies for predicting SOC fractions remains unclear, particularly in forest soils where environmental variability is high. Moreover, the spectral mechanisms underlying model predictions are rarely examined, limiting the ability to determine whether machine learning models capture physically meaningful soil signals.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

To address these gaps, this study systematically evaluates the potential of Vis–NIR spectroscopy to predict forest SOC fractions using both laboratory and in-situ measurements. Specifically, we compare multiple modeling strategies, including laboratory calibration, laboratory-to-field model transfer, spectral-correction-assisted transfer, and direct in-situ modeling, under varying environmental conditions. The specific objectives of this study were to: (1) evaluate the feasibility of predicting forest SOC fractions (POC and MAOC) using in-situ Vis–NIR spectroscopy; (2) compare the effectiveness of different modeling strategies for field-based spectral prediction; (3) identify the dominant spectral regions contributing to model predictions and examine their consistency with known soil mineral–organic interaction mechanisms; and (4) assess the robustness of spectral models under environmental variability using independent cross-year validation.

2. Materials and methods

2.1. Study area and soil samples

The study was conducted on the northern slope of Changbai Mountain, within the Changbai Mountain National Nature Reserve (41°41′49″-42°51′18″ N, 127°42′55″-128°16′48″ E), Jilin Province, Northeast China (**Fig. 1a, b**). The area lies at an average elevation of 758 m and experiences a typical East Asian continental monsoon climate, with mean annual temperatures of 2.5-5.5°C, 700-900 mm of precipitation, and roughly 2300 hours of sunshine per year. Soils are predominantly albic dark brown soils developed on basalt, characterized by a humus-rich dark surface layer and a grayish albic subsurface horizon.

Soil sampling was conducted during two independent field campaigns in 2024 and 2025 across seven representative forest sites (**Fig. 1c**). The same sampling protocol and sample preparation procedures were applied in both years. The spatial arrangement of sampling points differed between the campaigns due to

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

field constraints, as described below. In 2024, three forest sites were sampled. At each site, nine 2 × 2 m quadrats were established, and three sampling points were randomly selected within each quadrat. In 2025, four forest sites were sampled. At each site, five 2 × 2 m quadrats were established, with four sampling points randomly selected per quadrat. At each sampling point, two adjacent soil cores were collected and combined to form a composite sample. Soil samples were taken from four depth intervals (0-10, 10-20, 20-30, and 30-40 cm), yielding ~1 kg per composite sample. A total of 624 soil samples were collected across two sampling campaigns: 324 in 2024 and 300 in 2025. After removing visible roots and coarse debris, samples were placed in labeled plastic bags, air-dried, gently crushed, and sieved through a 2 mm mesh prior to laboratory analyses and spectral measurements.

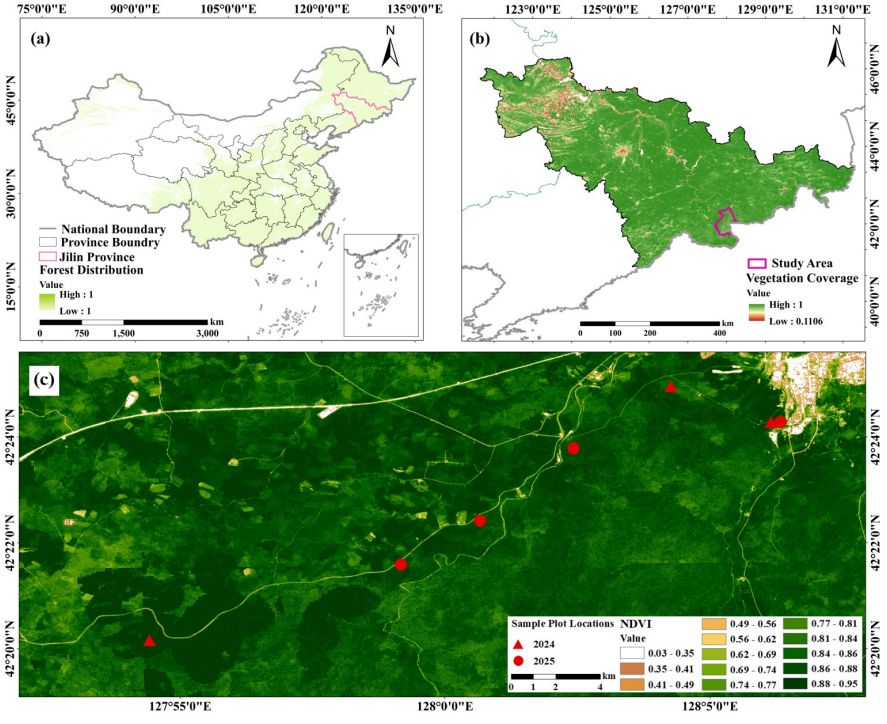

Figure 1. The location of the study area and soil sampling sites.

2.2 Laboratory measurement of SOC fractions

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

SOC was fractionated into particulate organic carbon (POC, ≥53 μm) and mineral-associated organic carbon (MAOC, <53 μm) using a size fractionation methodology similar to that outlined in Cotrufo et al. (2019). Briefly, 10 g of air-dried soil was mixed with 50 mL of 5.0 g L^{-1} sodium hexametaphosphate by manual shaking for 5 min, then oscillated at 90 rpm for 18 h. The suspension was then washed through a 53 μm sieve until the resulting filtrate was clear. The fraction retained on top of the sieve (e.g., POM, >53 μm) and the fraction which passed through the sieve (e.g., MAOM, <53 μm) were collected in pre-weighed aluminum containers, dried at 60°C for at least 48 h. After drying, each fraction was gently ground with a mechanical grinder and sieved through a 0.15-mm mesh. Material retained on the sieve was re-ground until complete passage was achieved, ensuring sample homogeneity prior to elemental analysis. The carbon content for each isolated fraction was determined using an elemental analyzer (Vario MACRO cube).

2.3 Spectral data acquisition and analysis

2.3.1 Laboratory and in-situ spectral measurements

Both in-situ and laboratory spectra were acquired using an ASD FieldSpec4 spectrometer (Analytical Spectral Devices, Inc.) equipped with a high-intensity contact probe, covering the 350-2500 nm range at 1 nm resolution. The spectrometer was calibrated before each measurement using a white Spectralon® panel.

For in-situ measurements, soil cores from the four depth intervals were measured immediately after extraction. Care was taken to avoid gravel, roots, litter, and other debris. Five replicate spectra per layer were averaged to reduce surface heterogeneity. This approach preserved vertical soil stratification and minimized the lag between sampling and measurement.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

Laboratory spectra were obtained from air-dried, ground, and sieved samples placed in 8 cm diameter containers with leveled surfaces. For each sample, five replicate spectra were measured and averaged to produce a single representative spectrum. Consequently, 624 averaged spectra were obtained for in-situ measurements and 624 for laboratory measurements, resulting in 1248 spectra used for analysis.

2.3.2 Spectral morphology analysis

To provide a physically grounded interpretation of model outputs and to assess whether statistically significant wavelengths correspond to meaningful spectral features, spectral morphology was examined alongside machine-learning results. Spectra were grouped by soil horizon and averaged for each depth. The resulting depth-specific mean spectra from laboratory and in-situ measurements were compared, yielding eight representative spectral curves (**Fig. 4a**). The Morphological Interpretation of Reflectance Spectra (MIRS; Dematt$\hat{e}$ et al., 2014) framework was applied to extract three key attributes: (1) absolute reflectance, (2) spectral shape (slope and curvature), and (3) absorption features (depth, width, and area). These morphological descriptors were used to interpret the influential spectral regions identified by the models and to link them to soil properties.

2.4 Model development, transfer, and validation

Model development and evaluation were conducted using a stepwise framework under laboratory and in-situ conditions. The workflow consisted of: (1) laboratory-based model development and evaluation, in which six preprocessing methods were combined with twelve machine learning algorithms (72 combinations) to identify robust approaches; (2) direct transfer of laboratory-calibrated models to in-situ spectra (Workflow A); (3) transfer of laboratory models to EPO-corrected in-situ spectra (Workflow

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

B); (4) direct in-situ model development using field-acquired spectra (Workflow C); and (5) EPO-corrected in-situ model development (Workflow D) (**Fig. 2**). The laboratory-based modeling step served as the baseline for evaluating the transferability and applicability of the four subsequent in-situ application workflows. Based on laboratory performance and stability, the three best-performing models were selected for subsequent in-situ analyses. Model robustness was further evaluated through independent cross-year validation.

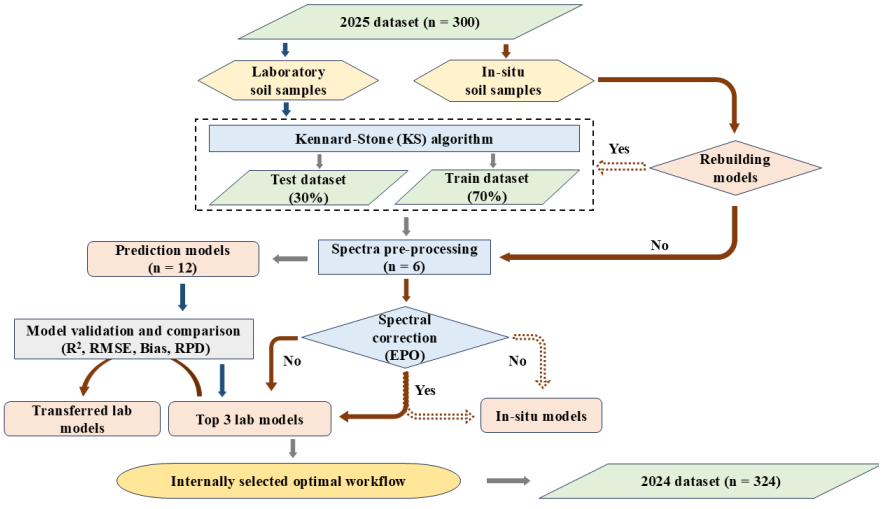

Figure 2. Flowchart of the modeling framework. Gray arrows: Shared Steps; Red arrows: In-situ-specific Steps (Dashed arrows distinguish laboratory-to-field model transfer from direct in-situ model development); Blue arrows: Lab-specific Steps.

2.4.1 Dataset partitioning strategy

The train and test datasets were partitioned using the Kennard-Stone (KS) algorithm (Kennard and Stone, 1969) based on laboratory spectra to ensure a representative coverage of the spectral space. To avoid potential data leakage arising from within-site correlations, the KS algorithm was implemented at the sampling-location level rather than at the individual-sample level. Specifically, all depth layers

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

corresponding to the same sampling location were assigned collectively to either the calibration or test set. This grouping strategy ensured strict independence between datasets by preventing samples from the same site from being split across calibration and validation subsets. This design was adopted to maximize spectral diversity within each subset and to support the primary objective of this study, which focuses on assessing the transferability of laboratory-calibrated models to in-situ spectral measurements under varying environmental conditions, rather than on strict spatial or site-level extrapolation.

2.4.2 Spectral preprocessing

Spectral preprocessing was applied to reduce noise, correct scattering effects, and highlight chemically relevant signals. Different preprocessing options were considered. The reflectance spectra (REF) were first truncated to 400-2500 nm to remove the noisier parts of the spectra. Reflectance data were then converted to absorbance ($AB = \log_{10}(1/REF)$) to improve linearity between spectral response and soil constituent concentrations. Additional preprocessing included first derivative (FD) and standard normal variate (SNV), as well as their combinations with AB (AB_FD, AB_SNV). FD was calculated using a simple finite difference approach (i.e., first-order difference between adjacent wavelengths) without additional smoothing, whereas SNV corrected for scattering and intensity variations (Wang et al., 2022). The influence of different preprocessing approaches on model stability and transferability, assessed using internal and external validation sets, was evaluated during model development, and the strategy yielding the most robust performance was ultimately retained.

2.4.3 Laboratory model development

To assess the predictive performance of different modeling approaches under controlled laboratory conditions and to select models suitable for in-situ application, twelve widely used spectral algorithms

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

were evaluated using laboratory spectra. These models cover a broad range of linear and nonlinear approaches commonly applied in soil spectroscopy studies. Linear models included Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), and Lasso regression. Nonlinear approaches comprised memory-based learning (MBL), tree-based methods (Cubist, Random Forest, Gradient Boosting Decision Trees), kernel-based methods (SVMR, GPR), and neural networks (ANN, RNN, CNN). All models were trained and optimized using five-fold cross-validation on the 2025 training set. Model training and evaluation were implemented in R (v4.4.3) and Python (v3.9.0) using standard packages for modeling and machine learning. The detailed descriptions of the models and their hyperparameters are provided in Supplementary Material S1.

2.4.4 Spectral correction method

In-situ soil spectral measurements are strongly influenced by environmental factors, particularly soil moisture, which may obscure spectral features associated with intrinsic soil properties and degrade predictive performance (Roudier et al., 2017). To harmonize measurement conditions between laboratory and in-situ spectra, the in-situ Vis–NIR spectra were corrected using the EPO method (Minasny et al., 2011).

Paired laboratory and in-situ spectra were constructed by matching samples collected from identical sampling locations and depth layers. For each paired observation, a difference spectrum (laboratory minus in-situ) was calculated to characterize systematic measurement-induced variability. The matrix of difference spectra was subjected to principal component analysis (PCA), and the resulting eigenvectors were used to define an orthogonal subspace representing external parameter effects. A projection matrix was subsequently constructed to remove this subspace from the in-situ spectra prior to model transfer or

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

in-situ model calibration.

The number of principal components retained (c) was determined as the minimum number required to explain 95% of the variance in the difference spectra. This threshold was selected as a conservative compromise to remove dominant measurement-related variability while minimizing the risk of eliminating soil-intrinsic spectral information. The derived projection matrix was consistently applied in all subsequent analyses.

2.4.5 Transfer of laboratory models to in-situ spectra

To evaluate the transferability of laboratory-calibrated models under field conditions, the three best-performing models identified in the laboratory evaluation were selected for subsequent in-situ analyses (**Fig. S4**). The laboratory-calibrated models served as the baseline reference for all transfer analyses. As Workflow A, laboratory-calibrated models were directly applied to in-situ spectra without recalibration or spectral correction. This direct transfer reflects the intrinsic ability of spectral–property relationships learned under controlled laboratory conditions to generalize to field-acquired spectra. As Workflow B, in-situ spectra were first corrected using EPO (Minasny et al., 2011) before model transfer. The same EPO projection matrix derived from the training dataset was applied consistently across workflows.

2.4.6 In-situ spectral model development with and without spectral correction

Accordingly, direct in-situ spectral models were developed using field-acquired Vis–NIR spectra (Workflow C). The same three modeling algorithms identified from the laboratory evaluation were employed to ensure direct comparability with the transfer workflows. In addition, an EPO-assisted in-situ modeling workflow (Workflow D) was implemented, in which in-situ spectra were first corrected

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

using the same training-sample-derived EPO projection matrix before model calibration. For both Workflow C and Workflow D, model calibration, validation, and hyperparameter optimization strictly followed the procedures described above, ensuring methodological consistency and unbiased comparison across all workflows.

2.4.7 Cross-year environmental transfer evaluation

The dataset collected in 2025 served as the model development dataset and was used for calibration–test partitioning, cross-validation, hyperparameter optimization, spectral preprocessing parameterization, and EPO transformation matrix construction.

In contrast, the dataset collected in 2024 was kept fully independent of all model development procedures and used exclusively as an external validation dataset to evaluate model transferability and robustness across different environmental conditions. All preprocessing parameters, EPO projection matrices, and model hyperparameters were determined solely based on the 2025 data and subsequently fixed prior to prediction on the 2024 dataset.

Differences in spectral distributions between the 2024 and 2025 datasets are illustrated in **Fig. S2**, reflecting variations in soil moisture conditions and measurement environments between sampling campaigns.

2.5 Evaluation metrics

Model performance was assessed on both internal test sets and independent external validation sets using R^2, Root Mean Squared Error (RMSE), Bias, and the ratio of performance to deviation (RPD). During model development, a five-fold cross-validation (CV) procedure was applied to the training

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

dataset for model tuning and selection, and the optimal models were subsequently evaluated on the independent test and external validation datasets. The equations are as follows:

$$R^2 = 1 - \frac{\sum_i^n (\hat{y}_i - y_i)^2}{\sum_i^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_i^n (\hat{y}_i - y_i)^2}{n}} \quad (2)$$

$$Bias = \frac{\sum_i^n (y_i - \hat{y}_i)}{n} \quad (3)$$

$$RPD = \frac{SD}{RMSE} \quad (4)$$

where n is the number of samples, y_i and $\hat{y}_i$ are the measured and predicted values, respectively, $\bar{y}$ is the mean of measured values, and SD is the standard deviation of measured values. R^2 indicates the proportion of variance explained by the model, RMSE reflects the average prediction error, Bias reveals systematic over- or underestimation, and RPD evaluates predictive capability. RPD was used as a relative indicator of predictive performance with respect to data variability and was interpreted in a comparative rather than categorical manner (R.A. Viscarra Rossel et al., 2006).

2.6 Model interpretation analysis

Model interpretation was conducted using SHapley Additive exPlanations (Lundberg and Lee, 2017) to quantify the contribution of individual spectral bands to model predictions. SHAP has been widely applied in soil spectroscopy to identify informative wavelengths and to improve the interpretability of machine learning models (Padarian et al., 2020; Haghi et al., 2021), particularly when combined with spectral morphology analysis (Viscarra Rossel et al., 2022; Ma et al., 2023).In this research, SHAP was conducted exclusively for the best-performing model identified for each target variable. Given the fundamental structural differences between Cubist (global, rule-based modeling) and MBL (locally

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

adaptive modeling), distinct SHAP strategies were implemented to ensure consistency between the prediction mechanisms and the attribution analysis.

2.6.1 SHAP interpretation of Cubist models

Cubist operates under a single global prediction structure defined by rule-based regression functions. SHAP values were therefore computed directly from the fitted Cubist models to quantify wavelength-level contributions under a consistent global mapping between spectral inputs and target variables, enabling interpretation of global spectral importance patterns across the dataset.

2.6.2 Local SHAP interpretation for MBL models

In contrast to Cubist, MBL generates predictions using locally calibrated models that vary among target samples. Because standard SHAP implementations assume a fixed global prediction function, applying global SHAP directly to MBL outputs may yield ambiguous interpretations.

To ensure consistency between prediction and attribution, SHAP analysis for MBL was conducted at the local model level. For each test sample, a local calibration set was constructed by retrieving its k nearest neighbors in spectral space using Euclidean distance, consistent with the neighborhood definition and optimal k used in MBL prediction (Wold et al., 1998; Daelemans, 2009) A local PLSR model was fitted within each neighborhood, and SHAP values were computed relative to this local prediction function using a model-agnostic framework (Sokol and Flach, 2020). Ten test samples were selected via stratified sampling across the full target range to balance representativeness and computational feasibility. SHAP values were calculated separately for each local model and aggregated by averaging absolute contributions to derive stable spectral importance patterns.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

3. Results

3.1 Forest SOC fractions

The mean contents of POC and MAOC were 19.04 and 4.56 g kg^{-1} (**Fig.3**), respectively. Both POC and MAOC exhibited positively skewed distributions, with skewness values of 1.87 and 1.70, respectively, indicating the presence of high-value observations. The training and test sets showed comparable ranges (**Fig. 3**). For POC, values ranged from 1.28 to 134.88 g kg^{-1} in the training set and from 1.76 to 67.89 g kg^{-1} in the test set. For MAOC, the corresponding ranges were 0.48–17.62 g kg^{-1} and 1.05–11.47 g kg^{-1}, respectively (Fig. 3a). POC exhibited greater variability than MAOC across samples.

The relationships between SOC and its fractions were further examined using Pearson correlation analysis, showing a strong correlation between SOC and POC ($r = 0.99$) and a moderate correlation between SOC and MAOC ($r = 0.72$) (**Fig. S3**).

The distributions of POC and MAOC in the training and test sets were further compared using probability density functions (**Fig. 3b**). The density curves showed a high degree of overlap between the two subsets, indicating consistent distribution patterns and suggesting that the dataset partitioning is representative with minimal sampling bias.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

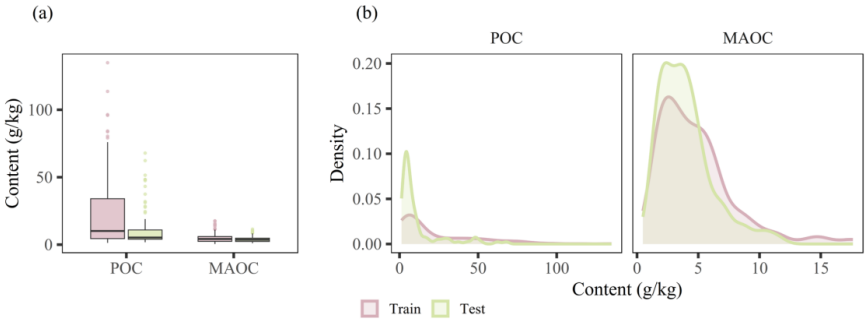

Figure 3. (a) Distributions of particulate organic carbon (POC) and mineral-associated organic carbon (MAOC) for the training (210 samples) and test (90 samples) sets; (b) probability density functions showing the distributions of POC and MAOC in the training and test sets.

3.2 Spectral characteristics of laboratory and in-situ soil measurements

The mean reflectance of in-situ soil samples was, on average, 71.45% lower than that of laboratory air-dried samples across the 400–2500 nm spectral range (**Fig. 4**). Reflectance reduction was particularly evident near 1400 nm and 1900 nm. Principal component analysis further revealed a clear separation between laboratory and in-situ spectra (**Fig. S1**), indicating a distinct shift in the spectral domain.

Across all soil layers, the raw reflectance spectra (**Fig. 4a**) exhibited a consistent three-stage pattern: increasing reflectance in the visible region (400–700 nm), a relatively stable plateau in the near-infrared region (700–1300 nm), and a gradual decline in the shortwave infrared region (1300–2500 nm). Under laboratory conditions, subsurface layers (10–40 cm) generally showed higher reflectance than the surface layer (0–10 cm), while the surface layer exhibited lower reflectance and smoother spectral concavity.

The second derivative of the Kubelka–Munk function (**Fig. 4b**) revealed distinct spectral features in the visible region (Demattê et al., 2014). A feature around 580 nm showed higher intensity in the surface layer, whereas features near 450 nm were more pronounced in subsurface layers. The feature

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

around 710 nm also displayed clear vertical variation, with higher intensity in the surface layer. In the shortwave infrared region (**Fig. 4c**), several diagnostic features were observed. A prominent feature near 2210 nm showed stronger intensity in subsurface layers, whereas a feature near 2180 nm was more pronounced in the surface layer. Additional features around 2310 nm also exhibited depth-dependent variation.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

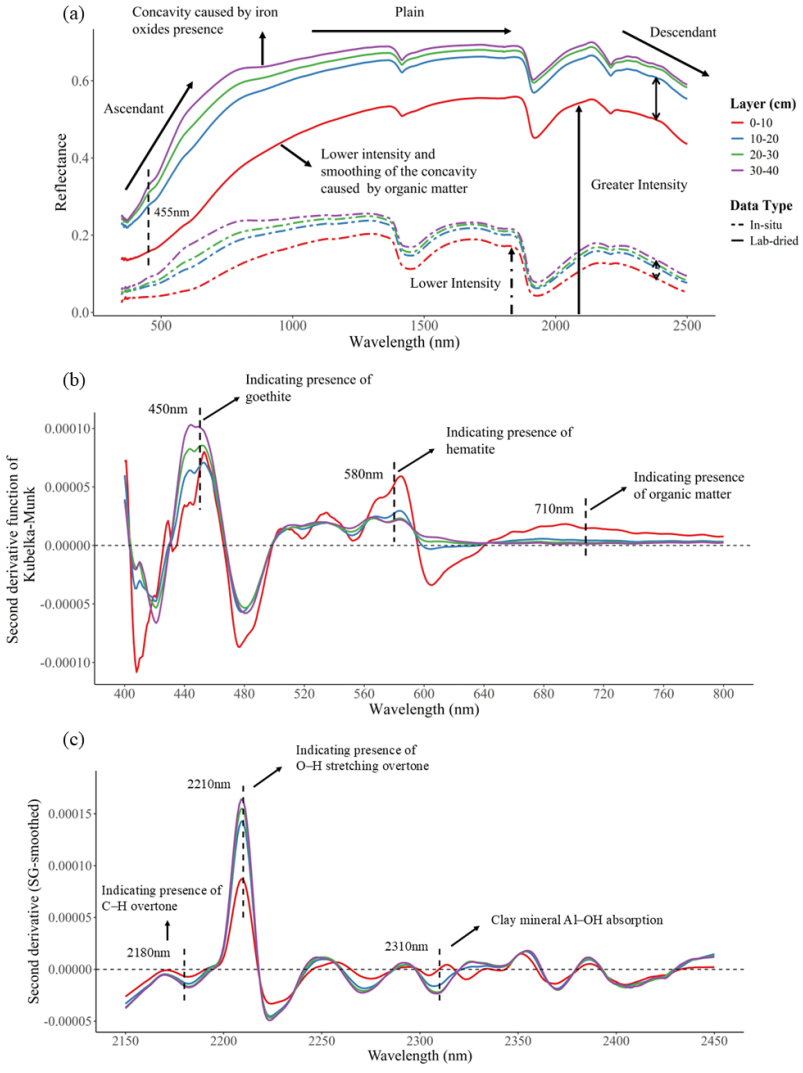

Figure 4. Spectral characteristics and feature analysis of soil samples. (a) Comparison of mean spectral reflectance (0–40 cm) between laboratory (solid lines) and in-situ (dashed lines) measurements; (b) Second derivative of Kubelka–Munk transformed laboratory spectra (400–800 nm); (c) Second derivative of Savitzky–Golay (SG) smoothed laboratory spectra (2150–2450 nm).

3.3 Laboratory predictions of SOC fractions

Among all evaluated models, PLSR, Cubist, and MBL consistently outperformed the others. For

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

POC prediction, R^2 values ranged from 0.85 to 0.92 for PLSR, 0.88 to 0.92 for Cubist, and 0.86 to 0.96 for MBL (**Fig. 5**). In contrast, MAOC prediction exhibited substantially lower accuracy, with R^2 values ranging from 0.30 to 0.65 for PLSR, 0.48 to 0.74 for Cubist, and 0.45 to 0.69 for MBL.

For both of the carbon fractions, the AB_SNV preprocessing consistently yielded the highest prediction accuracy. Under this unified preprocessing condition, POC achieved the best performance with MBL, with predictions yielding an R^2 of 0.96, RMSE of 3.05, Bias of 0.48, and RPD of 4.67. By comparison, under the same preprocessing conditions, MAOC prediction achieved its optimal performance with Cubist, reaching an R^2 of 0.74, an RMSE of 1.15, a Bias of -0.26, and an RPD of 1.90.

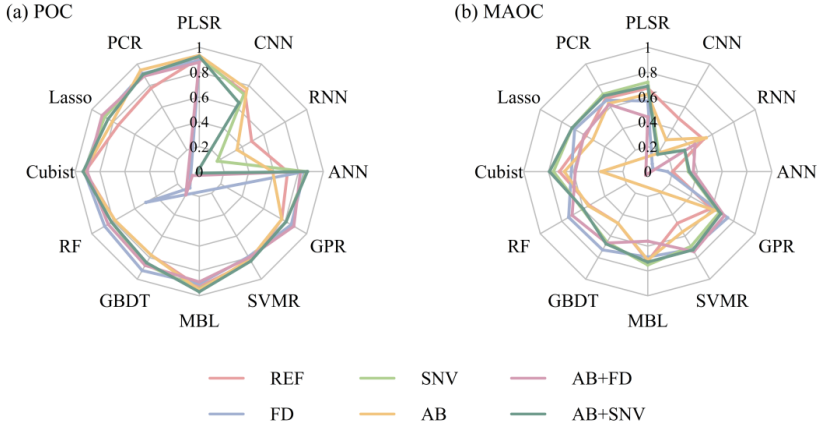

Figure 5. Comparison of predictive performance (R^2) of twelve machine learning models for particulate organic carbon (POC) and mineral-associated organic carbon (MAOC) based on laboratory Vis-NIR spectra. Radar charts illustrate model performance under different spectral preprocessing methods. The radius along each axis indicates the R^2 coefficient, with higher values indicating better predictive accuracy.

3.4 In-situ predictions of SOC fractions

Model transferability and in-situ applicability for POC and MAOC were assessed by comparing the predictive performance of the three best-performing laboratory models (**Fig. S4**) across the four in-situ

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

application workflows (A–D), including two laboratory-to-field transfer workflows (A–B) and two direct in-situ modeling workflows (C–D), with the optimal model–preprocessing combinations identified under each workflow (**Fig. 6**).

The direct application of laboratory-calibrated models to in-situ spectra (Workflow A) resulted in pronounced declines in predictive performance across all carbon fractions. POC showed a substantial decline in predictive performance following transfer, with R^2 decreasing from 0.85–0.96 under laboratory conditions to 0.04–0.70. In contrast, MAOC exhibited the weakest transfer performance, with R^2 declining from 0.30–0.74 to mostly below 0.43.

Workflow B partially mitigated the decline observed under Workflow A. For POC, R^2 values increased to 0.23–0.85, respectively, with reductions in RMSE relative to Workflow A. However, improvements varied across model–preprocessing combinations; for example, under SNV + PLSR, R^2 for POC decreased compared with Workflow A. MAOC prediction showed only marginal gains in R^2.

Workflow C consistently outperformed the transfer-based workflows. POC prediction improved substantially, with R^2 values ranging from 0.57 to 0.89. MAOC prediction also improved, with R^2 increasing to 0.50–0.71.

Calibrating models using EPO-corrected spectra (Workflow D) achieved performance comparable to or slightly better than Workflow C, although improvements were primarily observed for the Cubist model (POC: R^2 0.74–0.90; MAOC: R^2 0.53–0.66). For PLSR and MBL, EPO correction led to minor reductions or negligible changes in predictive accuracy compared with Workflow C.

For in-situ applications, Workflow D provided the most robust and stable performance for POC,

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

with the Cubist model and AB preprocessing yielding the best balance of accuracy and robustness. In contrast, MAOC prediction performed best under Workflow C with MBL and AB-SNV preprocessing, whereas EPO correction did not provide additional improvement. The optimal in-situ modeling strategy identified for each target variable was subsequently evaluated using the independent 2024 dataset.

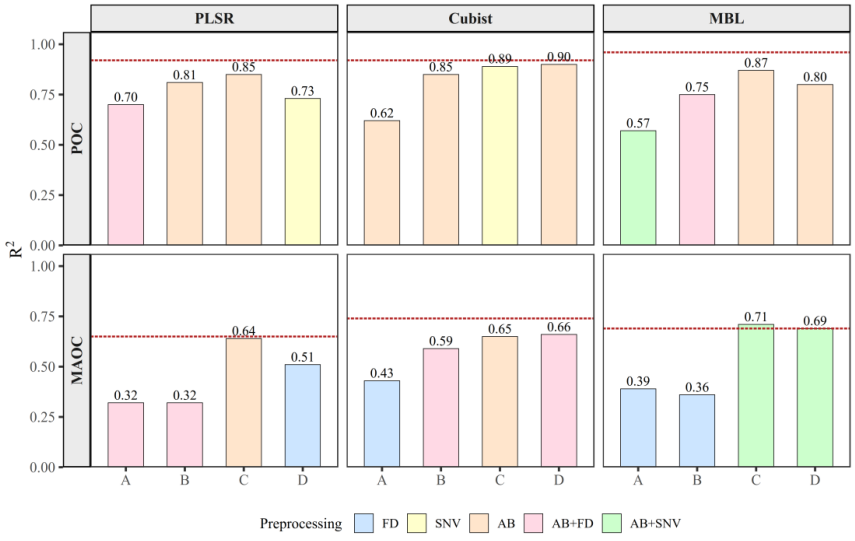

Figure 6. Performance of PLSR, Cubist, and MBL models across the four in-situ application workflows: (A) direct transfer of laboratory-calibrated models to in-situ spectra; (B) transfer of laboratory-calibrated models to EPO-corrected in-situ spectra; (C) direct in-situ model development; and (D) EPO-corrected in-situ model development. Acronyms in the legend denote preprocessing methods: REF, raw reflectance spectra; FD, first derivative; SNV, standard normal variate; and AB, absorbance transformation. The red dashed line indicates the optimal R^2 obtained under laboratory conditions.

3.5 Interpretation of the optimal SOC fractions prediction models

SHAP analysis was first conducted to interpret the best-performing laboratory models (models using spectra preprocessed with the AB_SNV transformation (**Fig. 7a**). The patterns of relative contributions for POC and MAOC were relatively stable across wavelengths, reflecting the consistent spectral responses observed under controlled laboratory conditions. For POC, the most influential wavelengths

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

were primarily concentrated between 2,200 and 2,400 nm, which are commonly associated with aliphatic C–H absorption features and mineral–organic matter interactions. In contrast, MAOC contributions were mainly located in the 500–600 nm region, likely related to iron oxide absorption features and aromatic organic matter components.

For the in-situ models, the optimal POC model was developed using EPO-corrected spectra. SHAP analysis indicates that EPO enhances chemically relevant spectral bands, resulting in wavelength contribution patterns that more closely resemble those observed under laboratory conditions (**Fig. 7b**). This suggests that EPO effectively suppresses environmental noise while preserving meaningful spectral information for POC prediction, which is consistent with its underlying. In contrast, the best MAOC model was constructed directly from uncorrected in-situ spectra and exhibited weak, localized SHAP responses across the spectral range. This indicates that environmental noise in field-acquired spectra strongly masks the characteristic spectral signals of MAOC, and that EPO correction may further remove weak but informative features. These results highlight the inherent difficulty of retrieving MAOC from in-situ hyperspectral observations.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

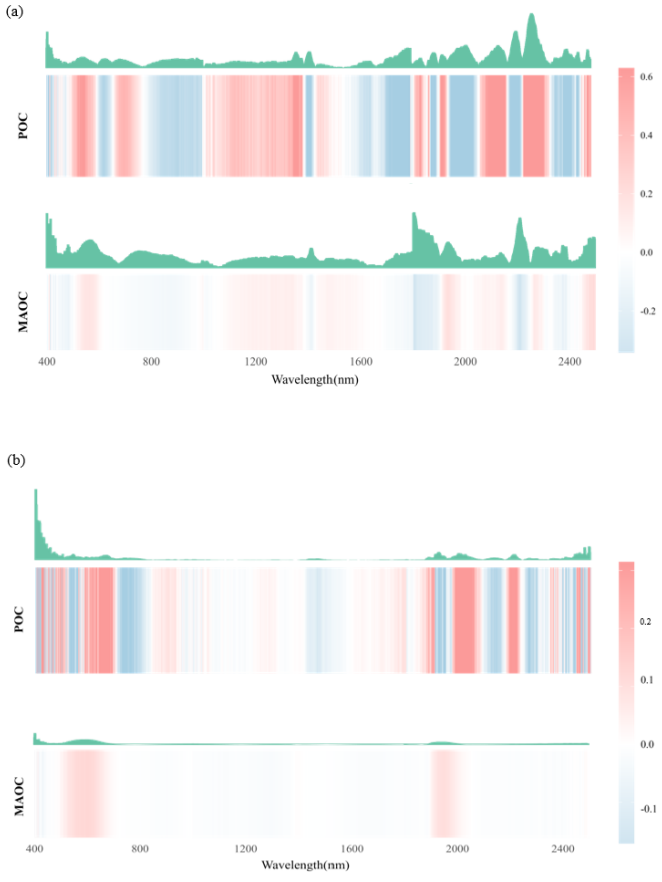

Figure 7. Feature importance for different kinds of organic carbon derived from the SHAP values of the optimal models. (a) Laboratory models; (b) In-situ models. The heatmap below each line graph shows SHAP-derived feature importance for stratified sampling. The bars above each heatmap represent absolute mean SHAP values across all samples.

3.6 Cross-year transferability of the prediction models

To evaluate model robustness across varying environmental states, an independent dataset collected in 2024 was used to externally validate models developed on the 2025 dataset. Because both datasets were acquired at the same site but under different field conditions, the cross-year comparison mainly reflects environmental effects, particularly differences in soil moisture and in-situ measurement states,

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

rather than true temporal dynamics. Therefore, this analysis focuses on environmental transferability driven by spectral domain shifts.

Overall, predictive performance declined on the independent 2024 dataset compared with the 2025 test dataset (**Fig. 8**), indicating reduced robustness under altered environmental conditions. For laboratory-based models, POC predictions retained relatively high accuracy on the external dataset (R^2 = 0.81, RPD = 1.88), suggesting moderate environmental transferability. In contrast, MAOC predictions showed substantial degradation, with an R^2 of 0.35 and RPD below 1.0, indicating very limited transferability for this fraction.

In-situ models exhibited consistently lower predictive performance on the external dataset. POC predictions declined to fair accuracy (R^2 = 0.62), whereas MAOC predictions remained unreliable (RPD < 1.0). Comparison of marginal density distributions (**Fig. 8**) further revealed that laboratory-based POC predictions exhibited greater overlap between the calibration and external datasets, whereas in-situ models showed pronounced distributional shifts, indicative of a stronger spectral-domain mismatch.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

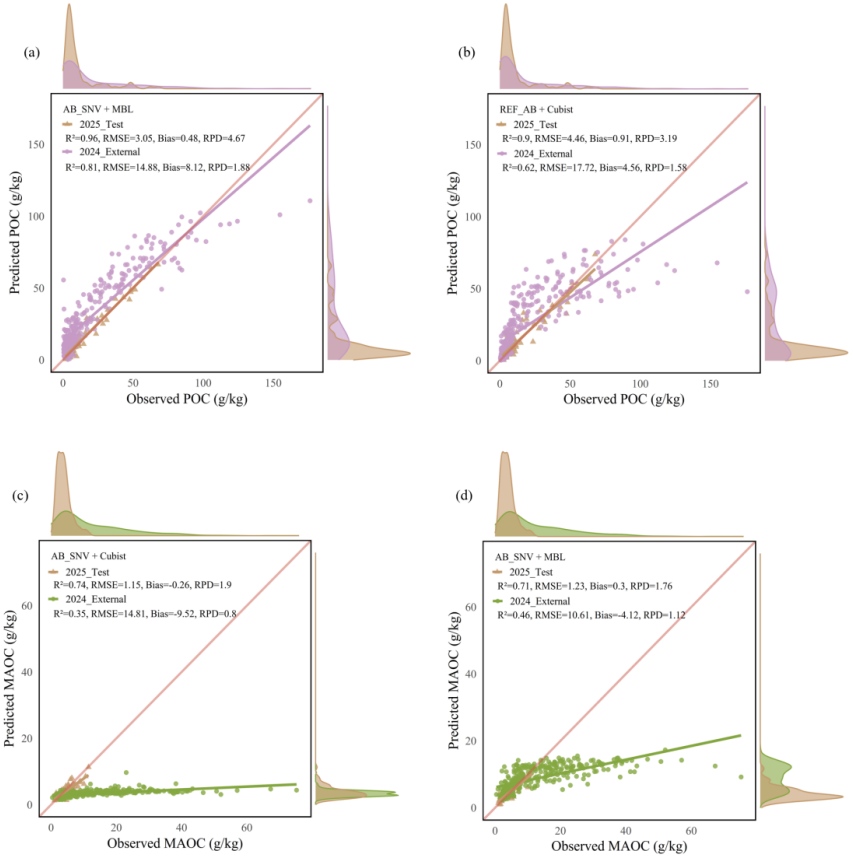

Figure 8. Comparison of environmental transferability between optimal laboratory and in-situ models. The left column shows the prediction performance of laboratory-based models, while the right column shows the prediction performance of in-situ models. Marginal density distributions of the test (2025) and external (2024) datasets are shown along the top and right axes.

4. Discussion

4.1 Feasibility of Vis-NIR prediction for SOC fractions

This study systematically evaluated 12 machine learning algorithms to predict forest soil organic carbon fractions from laboratory Vis-NIR spectra. Under controlled laboratory conditions, the models achieved reliable predictions for POC and MAOC (**Fig. 5**), confirming the feasibility of Vis-NIR

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

spectroscopy for estimating SOC fractions in forest soils. Similar levels of accuracy have been reported in previous studies, with R^2 values of 0.65 and 0.67 for POC and MAOC, respectively (Dai et al., 2024). Among the evaluated algorithms, PLSR, Cubist, and MBL consistently ranked among the best-performing models, indicating that both linear and nonlinear approaches can achieve high predictive accuracy when spectral variability is minimized and measurement conditions are well controlled.

However, previous studies have reported inconsistent conclusions regarding the optimal modeling strategy, with model performance varying across datasets, target fractions, and spectral configurations (Das et al. 2023; Munnaf and Mouazen, 2022). Together with our results, this variability suggests that no single algorithm can be universally recommended for Vis-NIR-based prediction of SOC fractions. Instead, systematic comparison of multiple modeling strategies remains necessary to identify models that best match the spectral characteristics and compositional properties of the target fractions.

When the best-performing laboratory models were directly applied to in-situ spectra, prediction accuracy declined markedly (**Fig. 6**). This reduction is consistent with the pronounced spectral discrepancies between laboratory and in-situ measurements (**Figs. 4a, S1**), where in-situ spectra exhibited lower reflectance due to soil moisture effects. Moisture-induced attenuation masked intrinsic absorption features, reduced the signal-to-noise ratio, and shifted the spectral feature space, thereby limiting the applicability of models trained under controlled laboratory conditions (Piccini et al., 2024). Under this transfer scenario, POC retained only moderate predictability, whereas MAOC prediction deteriorated substantially, reflecting its weaker and more indirect spectral expression.

By contrast, models developed directly from in-situ spectra achieved higher predictive accuracy than laboratory models applied to in-situ spectra. Because these models incorporate environmental

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

variability during calibration, they better capture both intrinsic spectral–chemical relationships and systematic field-induced spectral variation. These results highlight the importance of calibrating models directly under in-situ conditions when accurate field predictions are required, particularly for fractions such as MAOC that are sensitive to environmental noise.

4.2 Role of EPO correction across in-situ spectral application

EPO played a differentiated role across the in-situ spectral application scenarios examined in this study, highlighting both its advantages and limitations depending on the modeling strategy and target variable. For laboratory-to-field model transfer, EPO substantially improved the applicability of laboratory-calibrated models to in-situ spectral data. Compared with the direct application of laboratory models to raw in-situ spectra, the use of EPO-corrected spectra consistently resulted in higher predictive performance and lower prediction errors (**Fig. 6**), particularly for the PLSR and Cubist models. This improvement demonstrates the effectiveness of EPO in reducing spectral discrepancies between laboratory and field measurements.

By removing spectral components orthogonal to the target soil properties—such as variability due to soil moisture, surface roughness, and particle-size heterogeneity—EPO reduces non-informative spectral variance and improves the statistical compatibility between laboratory and in-situ spectral domains. As a result, models calibrated under controlled conditions become more transferable to field-acquired spectra.

In contrast, the influence of EPO on models calibrated directly with in-situ spectra was strongly model-dependent. For PLSR and MBL, EPO correction resulted in reduced predictive performance, suggesting that the orthogonalization process may inadvertently remove spectral information relevant to

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

soil carbon fractions. Both PLSR and MBL rely strongly on covariance structures and spectral similarity patterns, which may be partially disrupted when spectral components are removed during EPO correction. Conversely, Cubist models exhibited improved performance after EPO correction. The rule-based structure and local linear regression framework of Cubist allow it to exploit the simplified feature space generated by EPO while retaining relevant predictive information, explaining its stronger robustness under EPO-corrected in-situ modeling conditions.

These results indicate that the effectiveness of EPO depends strongly on the modeling objective and algorithmic structure. Rather than acting as a universally optimal preprocessing method, EPO functions as a conditional tool whose benefits are most evident when reducing measurement-condition mismatch between laboratory and in-situ spectra.

4.3 Spectral band contributions and mechanistic interpretation of SOC fractions

SHAP analysis of the optimal laboratory models showed that the spectral regions contributing to SOC fraction predictions align with known soil organic–mineral interactions (**Fig. 7a**). These contributions match the diagnostic spectral patterns identified in the visible and shortwave infrared regions (**Fig. 4**), indicating that the models capture physically meaningful relationships rather than purely statistical correlations.

For the POC, dominant contributions were concentrated in the shortwave infrared region, characterized by absorption features associated with organic functional groups. These signals are directly associated with plant-derived residues and microbial products, which constitute the main components of particulate organic matter (Dematt$\hat{e}$ et al., 2004; Stenberg et al., 2010). As a result, POC exhibits a relatively strong and direct spectral expression, enabling robust and stable model predictions across

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

different modeling scenarios.

In contrast, MAOC showed stronger contributions in the visible region, where spectral variability is largely controlled by mineralogical features. This reflects the fact that MAOC is primarily associated with mineral surfaces rather than existing as a spectrally distinct organic phase. Consequently, its spectral representation is mediated through mineral-related signals, which are not uniquely specific to MAOC but are influenced by broader soil compositional characteristics (Schwertmann and Taylor, 1989; Chen et al., 2014; Kleber et al., 2015).

The comparatively lower prediction accuracy of MAOC is therefore not attributable to a single factor but rather reflects the combined effects of its spectral representation and the data characteristics. First, the mineral-mediated nature of MAOC results in a less specific, more context-dependent spectral signal than POC. Second, the narrower concentration range and lower variability of MAOC observed in this study reduce the effective signal-to-noise ratio available for model calibration. Third, because mineral-associated spectral signals are particularly sensitive to environmental conditions such as soil moisture and surface heterogeneity (Dai, 2025), which can substantially alter spectral responses in the field (Roudier et al., 2017), MAOC predictions are more susceptible to spectral-domain shifts under in situ and cross-year conditions.

These factors highlight two fundamentally different modes of spectral detectability: a direct, organic-dominated signal for POC in the shortwave infrared region, and an indirect, mineral-mediated signal for MAOC in the visible region (Breure et al., 2025). This contrast is consistent with the conceptual basis of soil organic matter fractionation, in which size-based separation distinguishes functionally distinct carbon pools. POC represents a relatively labile pool dominated by plant residues and microbial

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

products, characterized by rapid turnover and strong intrinsic spectral signatures. In contrast, MAOC represents a more stable pool in which organic compounds are associated with mineral surfaces and protected from decomposition, resulting in a weaker and less specific spectral expression mediated by mineral-related signals (Wang et al., 2024; Yang et al., 2023).

These differences indicate that the observed variation in spectral contributions and model performance reflects not only spectroscopic behavior but also fundamental differences in the form and function of SOC fractions. This mechanistic distinction explains the contrasting predictive accuracy and transferability between POC and MAOC. It highlights the importance of considering both compositional characteristics and data structure when interpreting and applying Vis–NIR-based soil carbon models.

4.4 Challenges in the robustness of soil carbon models under cross-year environmental variability

The cross-year validation results highlight the challenges of achieving robust Vis–NIR-based prediction of soil carbon fractions under changing environmental conditions. The observed performance degradation is primarily attributable to environmental state–induced spectral domain shifts rather than to model instability alone (**Fig. S2**). Variations in soil moisture and associated changes in surface conditions substantially alter the statistical structure of in-situ spectra (Roudier et al., 2017), thereby limiting model generalization beyond the calibration domain.

Differences in robustness between laboratory-based and in-situ models reflect contrasting levels of spectral control. Laboratory spectra are obtained under standardized conditions, resulting in more stable spectral–chemical relationships and relatively stronger transferability for POC. In contrast, in-situ spectra integrate multiple sources of uncontrolled variability, including moisture fluctuations and surface

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

heterogeneity, which increase spectral domain mismatch when models are applied across environmental states. These findings demonstrate that high predictive accuracy in a single sampling campaign does not necessarily guarantee stable performance across years or conditions.

The influence of spectral correction further illustrates the conditional nature of model robustness. EPO improved laboratory-to-field transferability for POC by suppressing moisture-dominated spectral variance that is largely orthogonal to the target variable. However, the benefits of EPO were limited or even negative for MAOC. This likely reflects a violation of the orthogonality assumption underlying EPO: unlike POC, MAOC is closely associated with mineral surfaces and surface-bound water, resulting in statistical coupling between moisture-related spectral features and MAOC content.

SHAP analysis further supports this mechanistic interpretation by demonstrating that dominant spectral contributions captured by the models correspond to known soil mineralogical and organic interaction processes. Together, these results indicate that robust application of Vis–NIR spectroscopy for predicting forest SOC fractions requires explicit consideration of environment–driven spectral domain shifts and independent validation under contrasting field conditions.

From a methodological perspective, the optimal modeling strategy depends strongly on the intended application context. Laboratory-based models remain valuable for controlled spectral characterization and for establishing standardized spectral–chemical relationships. However, when the objective is reliable field prediction under environmentally variable conditions, direct in-situ calibration offers a more robust strategy because it incorporates environmental variability into model development.

The role of EPO should likewise be interpreted conditionally. EPO is most effective when the objective is to improve laboratory-to-field model transfer by reducing systematic spectral discrepancies,

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

particularly for spectrally intrinsic fractions such as POC. In contrast, when models are calibrated directly using in-situ spectra, the benefits of EPO become model- and target-dependent and may even be negative for proxy-mediated variables such as MAOC. Overall, these findings demonstrate that robust field spectroscopy depends not only on algorithm selection but also on aligning preprocessing and calibration strategies with the spectral expression of the target soil property.

5. Conclusion

This study evaluated the capability of Vis–NIR spectroscopy to predict SOC fractions in forest soils using either laboratory or in-situ spectra, multiple modeling strategies, spectral pre-processing options, and independent cross-year validation. Under laboratory conditions, reliable predictions were obtained for both POC and MAOC. The AB_SNV preprocessing strategy produced the most consistent results, with MBL performing best for POC and Cubist providing the most accurate predictions for MAOC.

Model performance declined substantially when laboratory-calibrated models were directly transferred to in-situ spectra, primarily due to environmental spectral discrepancies associated with soil moisture and surface conditions. This illustrates the importance of dedicated transfer strategies. Spectral correction using EPO improved laboratory-to-field model transfer for POC, but its effectiveness varied across algorithms and target fractions. In contrast, models calibrated directly using in-situ spectra achieved more stable predictions, indicating that incorporating environmental variability during calibration is essential for reliable field applications.

The interpretation of the calibrated models further demonstrated that they captured physically meaningful spectral information: POC predictions were mainly associated with organic absorption features in the shortwave infrared region, whereas MAOC predictions relied more strongly on mineral

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

proxy signals in the visible range.

Overall, this study demonstrates the feasibility of directly quantifying SOC fractions (POC and MAOC) in-situ using Vis–NIR spectroscopy under forest field conditions. Providing new methodological insights into soil spectroscopy under field conditions and establishing a foundation for advancing in situ monitoring of forest soil carbon dynamics.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

Declaration of competing interest

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

Author contributions

JL performed the measurements and analyzed the data; JL and YZ conceived and executed the research and wrote the manuscript draft; WF, LS, and YL gave suggestions about the approach and contributed extensively to the paper. AW, JW, and PR reviewed and edited the manuscript. All co-authors revised the paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFF1300501), the National Natural Science Foundation of China (32271873), the Youth Start-up Fund, Institute of Applied Ecology, Chinese Academy of Sciences, the China Postdoctoral Science Foundation (2023M743699), and the Program of China Scholarship Council (Grant No. 202504910271).

The authors thank the reviewers for their constructive comments, which have helped improve our manuscript.

Data availability

The datasets generated during the current study are available from the corresponding author upon reasonable request.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

References

Abramoff, R.Z., Guenet, B., Zhang, H., Georgiou, K., Xu, X., Viscarra Rossel, R.A., Yuan, W., Ciais, P., 2022. Improved global-scale predictions of soil carbon stocks with millennial version 2. Soil Biology and Biochemistry 164, 108466. https://doi.org/10.1016/j.soilbio.2021.108466

Breure, T.S., De Rosa, D., Panagos, P., Cotrufo, M.F., Jones, A., Lugato, E., 2025. Revisiting the soil carbon saturation concept to inform a risk index in european agricultural soils. Nat Commun 16, 2538. https://doi.org/10.1038/s41467-025-57355-y

Chen, C., Dynes, J.J., Wang, J., Sparks, D.L., 2014. Properties of fe-organic matter associations via coprecipitation versus adsorption. Environ. Sci. Technol. 48, 13751–13759. https://doi.org/10.1021/es503669u

Conforti, M., Matteucci, G., Buttafuoco, G., 2018. Using laboratory vis-NIR spectroscopy for monitoring some forest soil properties. J Soils Sediments 18, 1009–1019. https://doi.org/10.1007/s11368-017-1766-5

Cotrufo, M.F., Ranalli, M.G., Haddix, M.L., Six, J., Lugato, E., 2019. Soil carbon storage informed by particulate and mineral-associated organic matter. Nat. Geosci. 12, 989–994. https://doi.org/10.1038/s41561-019-0484-6

Daelemans, W., 2009. Memory-based language processing [WWW Document]. URL https://www.cambridge.org/core/books/memorybased-language-processing/8B3E1D0E5A2EA90E9DF509B67E410033 (accessed 4.27.25).

Dai, L., 2025. Prediction of soil organic carbon fractions in tropical cropland using a regional visible and near-infrared spectral library and machine learning.

Dai, L., Xue, J., Lu, R., Wang, Z., Chen, Z., Yu, Q., Shi, Z., Chen, S., 2024. In-situ prediction of soil organic carbon contents in wheat-rice rotation fields via visible near-infrared spectroscopy. Soil & Environmental Health 2, 100113. https://doi.org/10.1016/j.seh.2024.100113

Das, B., Chakraborty, D., Singh, V.K., Das, D., Sahoo, R.N., Aggarwal, P., Murgaokar, D., Mondal, B.P., 2023. Partial least square regression based machine learning models for soil organic carbon prediction using visible–near infrared spectroscopy. Geoderma Regional 33, e00628. https://doi.org/10.1016/j.geodrs.2023.e00628

De Vos, B., Cools, N., Ilvesniemi, H., Vesterdal, L., Vanguelova, E., Carnicelli, S., 2015. Benchmark values for forest soil carbon stocks in europe: Results from a large scale forest soil survey. Geoderma 251–252, 33–46. https://doi.org/10.1016/j.geoderma.2015.03.008

Delahaie, A.A., Cécillon, L., Stojanova, M., Abiven, S., Arbelet, P., Arrouays, D., Baudin, F., Bispo, A., Boulonne, L., Chenu, C., Heinonsalo, J., Jolivet, C., Karhu, K., Martin, M., Pacini, L., Poeplau, C., Ratié, C., Roudier, P., Saby, N.P.A., Savignac, F., Barré, P., 2024. Investigating the complementarity of thermal and physical soil organic carbon fractions. SOIL 10, 795–812. https://doi.org/10.5194/soil-10-795-2024

Demattê, J.A., Bellinaso, H., Romero, D., Fongaro, C., 2014. Morphological interpretation of reflectance spectrum (MIRS) using libraries looking towards soil classification. Scientia Agricola 71, 509–520. https://doi.org/10.1590/0103-9016-2013-0365

Demattê, J.A.M., Campos, R.C., Alves, M.C., Fiorio, P.R., Nanni, M.R., 2004. Visible–NIR reflectance: A new approach on soil evaluation. Geoderma 121, 95–112. https://doi.org/10.1016/j.geoderma.2003.09.012

Haghi, K.R., Pérez-Fernández, E., Robertson, J., 2021. Prediction of various soil properties for a national spatial dataset of scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. Geoderma 396, 115071. https://doi.org/10.1016/j.geoderma.2021.115071

Jaconi, A., Don, A., Freibauer, A., 2017. Prediction of soil organic carbon at the country scale: Stratification strategies for near-infrared data. European Journal of Soil Science 68, 919–929.

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

https://doi.org/10.1111/ejss.12485

Janik, L., Skjemstad, J., Shepherd, K., Spouncer, L., 2007. The prediction of soil carbon fractions using mid-infrared-partial least square analysis. Australian Journal of Soil Research 45, 73–81. https://doi.org/10.1071/SR06083

Kaiser, K., Guggenberger, G., 2003. Mineral surfaces and soil organic matter. European Journal of Soil Science 54, 219–236. https://doi.org/10.1046/j.1365-2389.2003.00544.x

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148. https://doi.org/10.1080/00401706.1969.10490666

Kleber, M., Eusterhues, K., Keiluweit, M., Mikutta, C., Mikutta, R., Nico, P.S., 2015. Chapter one - mineral–organic associations: Formation, properties, and relevance in soil environments, in: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 1–140. https://doi.org/10.1016/bs.agron.2014.10.005

Koch, M., Schodlok, M.C., Guggenberger, G., Stadler, S., 2021. Effects of water tension and surface roughness on soil hyperspectral reflectance. Geoderma 385, 114888. https://doi.org/10.1016/j.geoderma.2020.114888

Lavallee, J.M., Soong, J.L., Cotrufo, M.F., 2020. Conceptualizing soil organic matter into particulate and mineral-associated forms to address global change in the 21st century. Global Change Biology 26, 261–273. https://doi.org/10.1111/gcb.14859

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777.

Ma, Y., Minasny, B., Demattê, J.A.M., McBratney, A.B., 2023. Incorporating soil knowledge into machine-learning prediction of soil properties from soil spectra. European Journal of Soil Science 74, e13438. https://doi.org/10.1111/ejss.13438

Minasny, B., McBratney, A.B., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L., Joalland, S., 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma 167–168, 118–124. https://doi.org/10.1016/j.geoderma.2011.09.008

Munnaf, M.A., Mouazen, A.M., 2022. Removal of external influences from on-line vis-NIR spectra for predicting soil organic carbon using machine learning. CATENA 211, 106015. https://doi.org/10.1016/j.catena.2022.106015

Pacini, L., Schiedung, M., Stojanova, M., Roudier, P., Arbelet, P., Barré, P., Baudin, F., Cambou, A., Cécillon, L., Heinonsalo, J., Karhu, K., McNally, S., Omondiagbe, P., Poeplau, C., Saby, N.P.A., 2025. Predicting the proportion of centennially stable soil organic carbon using mid-infrared spectroscopy. Geoderma 462, 117536. https://doi.org/10.1016/j.geoderma.2025.117536

Padarian, J., McBratney, A.B., Minasny, B., 2020. Game theory interpretation of digital soil mapping convolutional neural networks. SOIL 6, 389–397. https://doi.org/10.5194/soil-6-389-2020

Piccini, C., Metzger, K., Debaene, G., Stenberg, B., Götzinger, S., Borůvka, L., Sandén, T., Bragazza, L., Liebisch, F., 2024. In-field soil spectroscopy in vis–NIR range for fast and reliable soil analysis: A review. European Journal of Soil Science 75, e13481. https://doi.org/10.1111/ejss.13481

Qi, M., Chen, S., Wei, Y., Zhou, H., Zhang, S., Wang, M., Zheng, J., Viscarra Rossel, R.A., Chang, J., Shi, Z., Luo, Z., 2024. Using visible-near infrared spectroscopy to estimate whole-profile soil organic carbon and its fractions. Soil & Environmental Health 2, 100100. https://doi.org/10.1016/j.seh.2024.100100

Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, Diffuse reflectance spectroscopy in soil science and land resource assessment 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

Roudier, P., Hedley, C. b., Ross, C. w., 2015. Prediction of volumetric soil organic carbon from field-moist intact soil cores. European Journal of Soil Science 66, 651–660. https://doi.org/10.1111/ejss.12259

Roudier, P., Hedley, C.B., Lobsey, C.R., Viscarra Rossel, R.A., Leroux, C., 2017. Evaluation of two methods to eliminate the effect of water from soil vis–NIR spectra for predictions of organic carbon. Geoderma 296, 98–107. https://doi.org/10.1016/j.geoderma.2017.02.014

Schwertmann, U., Taylor, R. m., 1989. Iron oxides, in: Minerals in Soil Environments. John Wiley & Sons, Ltd, pp. 379–438. https://doi.org/10.2136/sssabookser1.2ed.c8

Silvero, N.E.Q., Di Raimo, L.A.D.L., Pereira, G.S., Magalhães, L.P. de, Terra, F. da S., Dassan, M.A.A., Salazar, D.F.U., Demattê, J.A.M., 2020. Effects of water, organic matter, and iron forms in mid-IR spectra of soils: Assessments from laboratory to satellite-simulated data. Geoderma 375, 114480. https://doi.org/10.1016/j.geoderma.2020.114480

Sokol, K., Flach, P., 2020. LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. https://doi.org/10.48550/arXiv.2005.01427

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter five - visible and near infrared spectroscopy in soil science, in: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., Shen, Z., 2022. Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. European Journal of Soil Science 73, e13271. https://doi.org/10.1111/ejss.13271

Viscarra Rossel, R.A., McGlynn, R.N., McBratney, A.B., 2006. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. Geoderma 137, 70–82. https://doi.org/10.1016/j.geoderma.2006.07.004

Viscarra Rossel, R. A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

Viscarra Rossel, R.A., Zhang, M., Behrens, T., Webster, R., 2024. A warming climate will make australian soil a net emitter of atmospheric CO2. npj Clim Atmos Sci 7, 1–11. https://doi.org/10.1038/s41612-024-00619-z

Wang, H.-P., Chen, P., Dai, J.-W., Liu, D., Li, J.-Y., Xu, Y.-P., Chu, X.-L., 2022. Recent advances of chemometric calibration methods in modern spectroscopy: Algorithms, strategy, and related issues. TrAC Trends in Analytical Chemistry 153, 116648. https://doi.org/10.1016/j.trac.2022.116648

Wang, Z., Chen, S., Lu, R., Zhang, X., Ma, Y., Shi, Z., 2024. Non-linear memory-based learning for predicting soil properties using a regional vis-NIR spectral library. Geoderma 441, 116752. https://doi.org/10.1016/j.geoderma.2023.116752

Wold, S., Antti, H., Lindgren, F., Öhman, J., 1998. Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems 44, 175–185. https://doi.org/10.1016/S0169-

https://doi.org/10.5194/egusphere-2026-2666
Preprint. Discussion started: 26 May 2026

7439(98)00109-9

Wu, Y., Peng, Z., Wang, X., Huang, J., Yang, L., Liu, L., 2025. Warmer climate reduces the carbon storage, stability and saturation levels in forest soils. Earth's Future 13, e2024EF004988. https://doi.org/10.1029/2024EF004988

Yang Y., Wu F., Wu Q., Zhu J., Ni X., 2023. Soil organic carbon associated with iron oxides in terrestrial ecosystems: Content, distribution and control. Chin. Sci. Bull. 68, 695–704. https://doi.org/10.1360/TB-2022-0728

Zhou, Z., Ren, C., Wang, C., Delgado-Baquerizo, M., Luo, Y., Luo, Z., Du, Z., Zhu, B., Yang, Y., Jiao, S., Zhao, F., Cai, A., Yang, G., Wei, G., 2024. Global turnover of soil mineral-associated and particulate organic carbon. Nat Commun 15, 5329. https://doi.org/10.1038/s41467-024-49743-7