Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Vis-NIR and MIR Sensor Synergy enables Robust Assessment of Spatio-Temporal Soil Property Alterations

Julia Pöhlitz

julia.poehlitz@geo.uni-halle.de

Martin Luther University Halle-Wittenberg

Talha Mahmood

Martin Luther University Halle-Wittenberg

Michael Denk

Anhalt University of Applied Science

Christopher Conrad

Martin Luther University Halle-Wittenberg

Research Article

Keywords: soil spectroscopy, multi-sensor synergy, spatial variability, temporal and vertical transferability, long-term soil monitoring

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11278430/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Vis-NIR and MIR Sensor Synergy enables Robust Assessment of Spatio-Temporal Soil Property Alterations

Julia Pöhlitz[a] | Talha Mahmood[a] | Michael Denk[b] | Christopher Conrad[a]

Institutional affiliations

[a] Department of Geoecology, Institute of Geosciences and Geography, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany.

[b] Digital Technologies in Plant Production, Department of Agriculture, Ecotrophology, and Landscape Development, Anhalt University of Applied Science, 06406 Bernburg, Germany.

Corresponding author

Julia Pöhlitz, Department of Geoecology, Institute of Geosciences and Geography, Martin Luther University Halle-Wittenberg, von-Seckendorff-Platz 4, 06120 Halle (Saale), Germany; Mail: julia.poehlitz@geo.uni-halle.de

Highlights

- Spatial organisation persisted over a decade; while properties changed to varying degrees.
- Vis-NIR and MIR synergy improved prediction, in property- and depth-dependent manner.
- $CaCO_3$ was best predicted and transferred most robustly across depth and time.
- Cubist achieved the highest mean predictive performance among the evaluated algorithms.
- Calibration needs differed by property; weak predicted properties gained little from more data.

Abstract

Understanding how intensive agricultural management affects the spatial organisation and temporal stability of soil properties is important for long-term experimentation and digital soil monitoring. We investigated changes between 2011 and 2021 at the 15-ha Crop Research Unit Merbitz (Central Germany) and assessed the performance and transferability of Vis-NIR, MIR and synergistic spectroscopy for predicting seven soil properties. Samples were collected on a 25-m grid (n = 480 per year). PLS, Cubist and Random Forest were evaluated with algorithm-specific feature-selection methods, varying calibration proportions, and vertical (topsoil-subsoil) and temporal transfer.

Field-scale spatial organisation remained broadly persistent over the decade, while the individual soil properties changed to different degrees. In particular, topsoil SOM showed an increased relative contribution of small-scale or unstructured variability, whereas other properties retained more persistent spatial patterns. Spectra were more similar between years in the topsoil than in the subsoil, although both depths showed substantial temporal continuity

Predictive performance was dependent on the spectral dataset, property and depth. The synergistic dataset significantly outperformed MIR in both depths and outperformed Vis-NIR in the topsoil. $CaCO_3$ was predicted best (RPD up to 5.54), whereas SOM and topsoil texture fractions were difficult to predict. Cubist achieved the highest mean predictive performance, and feature selection maintained comparable performance with fewer wavelengths. Gains from larger calibration proportion were property-dependent and limited for weakly predicted properties.

$CaCO_3$ retained comparatively high performance under both vertical and temporal transfer, whereas SOM, pH_{H2O} and texture fractions were more sensitive to transfer. These results indicate that persistent spatial and spectral characteristics can support, but do not guarantee transferability, which depends on the stability of the spectral-property relationship. Overall, the findings support property-specific calibration and transferability assessment as important components of efficient long-term spectroscopy monitoring.

Keywords: soil spectroscopy; multi-sensor synergy; spatial variability; temporal and vertical transferability; long-term soil monitoring

Introduction

Agricultural landscapes exhibit pronounced spatial and temporal variability driven by interacting pedogenic factors and anthropogenic management (Amundson & Jenny, 1997; Jenny, 1941; Tóth et al., 2018). In long-term experimental sites, land-use practices strongly influence both the magnitude and spatial organization of soil properties. Fine-scale variability in key soil characteristics such as pH-value, soil organic matter (SOM), and carbonate ($CaCO_3$) content is therefore critical for interpreting treatment effects and ensuring agronomic productivity (Cambardella & Karlen, 1999; Tilse et al., 2022). Crucially, robust digital soil mapping has become indispensable not merely for static soil characterization, but as an operational foundation for modern land system science (Minasny & McBratney, 2016; Verburg et al., 2015). In the context of global environmental change, targeting and planning sustainable landscape transformations such as crop diversification strategies and nature-based solutions, such as high-resolution soil organic carbon accounting and adaptive management, fundamentally requires a temporally stable, spatially explicit, and cost-efficient biophysical baseline of the underlying soil matrix.

However, many experimental fields retain persistent legacy effects from historical land use, former plot boundaries, and past fertilization regimes. These legacies can introduce structured variability that masks subtle treatment signals and complicates statistical inference (Cambardella & Karlen, 1999; Kuzyakov & Zamanian, 2019). To counteract this, intensive management strategies aimed at soil homogenization are commonly implemented. Practices such as repeated tillage and standardized fertilization are intended to reduce spatial heterogeneity and establish a uniform field baseline. Yet, the extent to which long-term management homogenizes soils is often assumed rather than empirically demonstrated. Moreover, tillage may reduce variability at certain scales while maintaining or even enhancing it at others, depending on soil structure and organic matter dynamics (Six et al., 2000). Evaluating the extent to which long-term management reduces spatial heterogeneity and modifies persistent spatial patterns therefore remains essential for robust experimental design and digital soil mapping.

Advances in soil spectroscopy offer new opportunities to address these challenges. Proximal sensing using reflectance spectroscopy has emerged as a rapid and cost-efficient alternative to conventional laboratory analyses (Stenberg et al., 2010; Viscarra Rossel et al., 2006). Visible and near-infrared (Vis-NIR, 350–2500 nm) spectroscopy enables the simultaneous estimation of multiple soil properties by capturing overtones and combination bands of molecular vibrations (Soriano-Disla et al., 2014; Stenberg et al., 2010). This spectral range is particularly sensitive to iron oxides, clay minerals, and organic matter components, but overlapping absorption features can limit predictive accuracy in complex soil matrices (Soriano-Disla et al., 2014; Stenberg et al., 2010).

Mid-infrared (MIR) spectroscopy addresses some of these limitations by capturing fundamental molecular vibrations, resulting in sharper and more diagnostic absorption features (Janik et al., 1998; Soriano-Disla et al., 2014; Viscarra Rossel et al., 2006). Comparative studies have shown that MIR often performs better than Vis-NIR for properties such as soil carbon, carbonate, and several textural variables, although performance depends on the target property and the sample set (Soriano-Disla et al., 2014; Viscarra Rossel et al., 2006). Nevertheless, the predictive strength of both Vis-NIR and MIR varies depending on soil composition, mineralogy, and concentration ranges.

While sensor synergy has been widely demonstrated to improve spatial prediction accuracy (Nawar et al., 2016; Viscarra Rossel et al., 2006), its capacity to maintain predictive robustness against decadal temporal drift and systematic environmental shifts remains largely unexplored. Previous applications have predominantly focused on spatial validation or short-term calibration transfer, leaving a critical knowledge gap regarding how dual-range sensor synergies perform when tracking multi-year soil property alterations across distinct profile depths.

An often-overlooked dimension in soil monitoring is vertical stratification. In agricultural systems, topsoil layers (0–30 cm) are highly dynamic due to intensive management, including tillage, fertilization, and residue incorporation. These processes promote partial homogenization but also introduce temporal variability. In contrast, subsoil layers (30–60 cm) are less affected by short-term management and can preserve intrinsic site characteristics. Consequently, the subsoil functions as a highly intact and persistent pedological archive that encapsulates long-term site evolutionary pathways and inherent geological properties rather than anthropogenic noise (Malone et al., 2009). Disentangling this vertical differentiation is not only crucial for developing transferable spectroscopic models, but also for evaluating long-term landscape-scale sustainability and soil functional resilience under shifting cultivation patterns. A key question is whether models calibrated on one depth interval can be reliably applied to another, particularly given depth-independent shifts in property ranges and sharp vertical gradients (such as the exponential decline in soil organic carbon). While topsoil sampling is relatively accessible, subsoil characterization remains labour-intensive and costly. Improving model transferability across depths would therefore enhance monitoring efficiency.

Beyond spatial and vertical variability, temporal dynamics pose a major challenge for digital soil mapping. Most spectroscopic models are calibrated on static datasets, yet long-term monitoring requires accounting for temporal changes specifically, whether multi-year shifts in soil properties generate novel spectral variations beyond the initial calibration space, alongside potential variations in instrumental conditions (Nocita et al., 2015; Sanderman

et al., 2023). Evaluating model stability over extended periods, such as a decade, is therefore critical to determine whether existing models remain predictive or if periodic updating and additional sampling becomes necessary.

Closely related is the issue of sampling efficiency. Robust model development typically relies on large calibration datasets, which conflicts with the need for cost-effective monitoring. Learning curve approaches, based on systematically reducing training sample size, allow quantification of how model performance scales with data availability (Ramirez-Lopez et al., 2013; Stevens et al., 2013). Multi-sensor combination has been shown to improve prediction accuracy and robustness (Schmidinger et al., 2024)). Whether this synergy advantage also reduces calibration sample size requirements remains unexplored. This is particularly relevant for scaling laboratory-based models to remote sensing applications, where ground reference data are limited (Semella et al., 2022). Ultimately, traditional, rigid sampling strategies such as high-density static grids or fixed transects hit severe economical and logistical limits when applied beyond experimental plots, as maintaining dense, repeated sampling campaigns over time is cost-prohibitive and physically disruptive to the soil matrix. Furthermore, because static architectures blindly allocate resources without accounting for local variance or model uncertainty, they fail to capture shifting process dynamics. Overcoming this requires transitioning toward model-driven adaptive sampling-designs (such as uncertainty-guided or covariate-conditioned sampling), where auxiliary sensor data dictate where sampling is genuinely required. Consequently, a profound mechanistic understanding of the temporal persistence of soil properties is urgently required to conceptualize and optimize these future, sensor-driven adaptive sampling frameworks.

Against this background, long-term agricultural experimental stations such as Merbitz provide a valuable framework for disentangling persistent site-specific soil patterns from management-related changes and ongoing pedogenic processes. Such sites are expected to retain a relatively stable pedological background structure, while repeated cultivation, contrasting experimental treatments, and changes in management history may progressively modify the spatial expression of more dynamic soil properties, particularly in the topsoil. At the same time, the heterogeneous arrangement of experiments and plot sizes provides a setting in which management-induced patterns may be superimposed on persistent natural soil variability rather than completely replacing it. We therefore expect the temporal development of soil properties to differ with depth: the topsoil should exhibit stronger signatures of management and mechanical homogenization, whereas the subsoil should retain a greater proportion of its pre-existing spatial and structural organization while continuing to undergo slower pedogenic change. This distinction provides a basis for assessing whether the underlying soil system remains sufficiently stable to support transferable spectral relationships across time and depth.

Based on these considerations, we hypothesize that (i) the spatial structure of relatively stable pedological properties will persist over the decadal observation period, while more dynamic soil properties will exhibit greater temporal variability; (ii) management-related changes will be more pronounced in the topsoil than in the subsoil, resulting in distinct depth-dependent patterns of temporal stability and spatial organization; and (iii) the persistence of the underlying soil structure will support the temporal and vertical transferability of Vis-NIR and MIR-based prediction models, although changes in the distribution and spatial organization of individual soil properties may affect model performance. Based on these hypotheses, this study aims to (i) quantify pedological shifts and soil homogenization over a decade, (ii) evaluate the temporal and vertical transferability of combined Vis-NIR and MIR models to assess whether spectral-analytical relationships remain robust against decadal property shifts at a given sampling location, and (iii) optimize sampling strategies through data reduction analysis as a methodological pathway toward adaptive landscape monitoring.

Material & Methods

Trial site

The experimental site for this study was located in Central Germany at the Crop Research Unit Merbitz. Situated in the southern part of Saxony-Anhalt, the site is at an elevation of approximately 160 meters above sea level and is part of the loess region known as the "Magdeburger Börde" (**Fig. 1**). As a highly representative loess-dominated landscape within the Central German Dry Region, the site features conditions characteristic of prime agricultural areas. While the regional loess substrate implies a baseline pedological uniformity, long-term experimental management and historical land-use legacies induce significant fine-scale spatial variability, making it an ideal proxy for studying landscape-scale agro-ecological transitions. The climate conditions at the research site in Merbitz are mainly influenced by its leeward position relative to the Harz Mountains. The area falls within the Central German Dry Region, characterized by average annual precipitation of approximately 430 mm and mean annual temperatures of 10.4 °C.

The general soil type is a Chernozem (FAO, 1998). The texture of the soil (0–60 cm) consists of 120 g kg^{-1} sand, 790 g kg^{-1} silt and 90 g kg^{-1} clay, constituting a silt loam (Gee and Bauder, 1986). The total organic carbon content is equal to 23 g kg^{-1} and the pH value is 7.5. The terrain features a consistent slope with a slight inclination of 0.5 ° running from northeast to southwest and forming a stagnic-Luvisol in the latter. The entire test site is under agricultural use.

The Crop Research Unit Merbitz is an important research facility known for conducting various experiments. It has been operational since 2012, conducting various experiments of different durations, sizes and research

questions across the 15-hectare experimental area. Following each experiment, a homogenization phase lasting several years is implemented before the commencement of the next experiment. To meet these requirements, and to track the efficacy of anthropogenic homogenization against inherent site heterogeneity, the derivation of various soil physical and chemical properties is necessary, including pH-value, plant available nutrients, $CaCO_3$ content, texture and SOM content.

Sampling design

Prior to the establishment of the Crop Research Unit Merbitz, the entire experimental site was sampled in fall 2011. In 2011 and 2021, soil samples were taken at regular intervals in a 25 x 25-meter grid across the entire site. This classic grid configuration was chosen to establish a baseline free of spatial structural bias, serving as a benchmark for evaluating conventional sampling efficiency against future sensor-driven, adaptive designs. The geographical coordinates of the sampling locations were determined using a Garmin GPS Map 62s. In each grid cell, soil samples were collected from topsoil (0–30 cm) and subsoil (30–60 cm). Five punctures were made with a geological drill (the so-called "Pürckhauer") in a Z-shaped pattern to obtain the samples, which were then combined into a mixed sample and on which the following measurements were carried out. A total of 240 samples were collected at each depth, resulting in 480 soil samples per year. The collected soil samples were air-dried at 40 °C, passed through a 2 mm sieve and divided into two parts for laboratory analysis and spectral measurements.

Conventional Laboratory Soil Analysis

To establish a robust reference baseline covering both primary mineral-chemical controls and physical matrix properties, soil samples were processed for standard analysis as follows. While additional agronomic and chemical parameters (such as total nitrogen and plant-available nutrients) were assessed as part of the broader Merbitz long-term monitoring program, this study deliberately focuses on the core variables that primarily govern the spectral response of the loess matrix in the Vis-NIR-MIR range, specifically organic carbon, mineral-bound water, $CaCO_3$ content and phyllosilicates (Soriano-Disla et al., 2014; Stenberg et al., 2010; Viscarra Rossel et al., 2006). These include particle size analysis, in which the determination of the dispersed sand, silt, and clay fractions (<2 mm), was conducted using a laser diffraction analyser (Helos/KR, Sympatec GmbH, Clausthal-Zellerfeld, Germany). The particle size data were categorized into eight subclasses, adhering to United States Department of Agriculture (USDA) guidelines, spanning the range of 0.5 to 2000 µm (Soil Science Division Staff, 2017).

SOM was determined using the loss-on-ignition method (DIN EN 15935). This involved weighing the soil sample before and after incineration in a muffle furnace at 550 ± 25 °C for 150 minutes to calculate the SOM content by mass difference.

Soil reaction, as indicated by pH_{H2O}, was determined using a glass electrode as well as electric conductivity (EC) at a 1-to-2.5 soil-to-solution (weight-to-volume) ratio according to standard guidelines (DIN e.V., 1994, 2021).

$CaCO_3$ content was assessed by releasing carbon dioxide (CO_2) through the addition of 10 % hydrochloric acid, using a Scheibler calcimeter (Carl Hamm, Essen, Germany) (ISO 10693). Therefore, the volume of carbon dioxide released after adding 4 M HCl (DIN e.V., 1995) was determined.

Laboratory Spectroscopic Analysis

For the spectral measurements, circular transparent Petri dishes were filled with each 45 g of the dried and homogenised soil samples. The samples were levelled carefully by shaking and manual adjustment to create a smooth surface without compacting the soil, minimizing the influence from surface roughness and particle size.

The Vis-NIR reflectance spectra (350–2500 nm) were acquired using an RS-3500 spectroradiometer (Spectral Evolution, Lawrence, MA, USA). The measurements were conducted in a dark, black-painted room to eliminate stray light. An external light source was positioned at a 45° angle to the sample. The fibre optic was placed at a nadir position 15 cm above the sample, resulting in a measurement diameter of approximately 7 cm.

Before measurements, the system was calibrated using a Zenith Polymer® Diffuse Reflectance Standard (~95% reflectivity). Calibration was repeated after every five samples. For each sample, twelve individual spectra were recorded (three measurements at four positions, rotating the dish by 90° each time) to account for surface inhomogeneities. These spectra were subsequently averaged within the R software environment to represent one mean spectrum per sample.

For MIR (Mid-Infrared) spectral measurement taken as contact measurements, an Agilent 4300 Handheld FTIR device was used, covering the spectral range from 650 cm^{-1} to 5000 cm^{-1} (2000 nm to 15,000 nm). The instrument featured a spectral resolution of 4 cm^{-1}. Three repeat measurements were recorded for each sample, slightly offsetting the contact probe (Diffuse Reflectance Interface) each time. A calibration using a Coarse Gold Reference Cap was performed every 15 minutes.

Spectral Data Processing & Modelling

Spectral Preprocessing & Analysis

1. *Vis-NIR Pre-Processing*

Vis-NIR spectra were trimmed to the 400–2450 nm range to remove noisy extreme edge regions. The spectral resolution was subsequently reduced to 5 nm resulting in 411 spectral bands. This reduction decreases redundancy

and collinearity among adjacent wavelengths while substantially lowering computational requirements with minimal loss of information, a standard approach in soil spectroscopy (Ng et al., 2019; Y. Zhou et al., 2022).

Spectral preprocessing was applied to mitigate the effects of light scattering, particle size variations, and baseline drift. We computed a Savitzky–Golay (SG) first derivative on the reduced spectral matrix (polynomial degree p = 2; window size, w = 21). Derivative transformations are widely preferred as they enhance subtle absorption features and diminish baseline effects while preserving spectral signatures (Rinnan et al., 2009; Savitzky & Golay, 1964). To avoid sequential smoothing artifacts, the SG derivative was applied directly to the resolution-reduced matrix. Subsequently, a standard normal variate (SNV) transformation was applied, as it yielded the best performance in our preliminary analysis. SNV is widely used in soil spectroscopy to minimize multiplicative scatter effects caused by physical sample properties (Rinnan et al., 2009).

2. MIR Pre-Processing

MIR spectral data were trimmed to the 650–4000 cm^{-1} range, which is standard in soil spectroscopy, as regions outside this interval are typically excluded from modelling due to their lower analytical relevance (FAO, 2024; Yang & Mouazen, 2012). The atmospheric CO_2 absorption bands (2268–2389 cm^{-1}) were excluded from analyses because atmospheric carbon dioxide contributes strong and highly variable features that may obscure soil-related information and introduce artefacts into calibration models (Baldacci et al., 1978).

MIR spectral data are characterized by high collinearity among adjacent wavenumbers due to broad overlapping absorption bands of soil organic matter, minerals, and moisture. Spectral density was reduced by retaining every fourth measured point after trimming, resulting in an effective resolution of approximately 15 cm^{-1} and 216 spectral bands. This subsampling retains the core spectral information while reducing redundancy and computational load, thereby increasing the model robustness (FAO, 2024; S. Li et al., 2025).

The effect of the pre-processing pipeline on the mean Vis-NIR and MIR spectra, shown separately for each depth (topsoil and subsoil) and survey year (2011 and 2021), is presented in **Fig. 2**. The transformed spectra display the characteristic absorption features of both spectral ranges following the SG first-derivative and SNV transformations. Overall, the 2011 and 2021 mean spectra converged closely across both sensors and depths.

3. *Synergistic Data*

Following spectral pre-processing, a synergistic dataset was generated by combining Vis–NIR and MIR data. A total of 411 Vis–NIR bands and 217 MIR bands were concatenated, resulting in 628 spectral bands in the synergy

dataset. This low-level concatenation strategy preserves all spectral information from both ranges while permitting each range to be pre-processed independently according to its own physical and instrumental characteristics (Barra et al., 2021). Crucially, this multi-sensor synergy is hypothesized to maximize model stability against anthropogenically induced changes in soil surface characteristics (e.g. via management or tillage drift), thereby serving as a reliable baseline to evaluate cross-scale digital soil mapping.

All pre-processing steps were applied identically to topsoil and subsoil samples and consistently across both survey years (2011 and 2021). This ensures that observed differences in model performance across soil depth, years, and spectral datasets reflect the true differences in predictive information content rather than artifacts of varying pre-processing workflows.

Calibration Size & Model Transferability

Three calibration transfer experiments were conducted to evaluate the predictive performance of spectral models under conditions of varying data availability, vertical depth extrapolation, and temporal domain shift. Each experiment was applied independently to each of the three spectral datasets and to each target soil property.

Experiment 1 – Variable Train/Test Splits

To quantify the sensitivity of model performance to calibration set size, models were trained and evaluated across a systematic range of train/test proportions from 10:90 to 90:10, in increments of 10%. For each split proportion, samples were partitioned using stratified random sampling, ensuring full range of measured values was represented in the calibration set. This approach generates a learning curve that characterizes how predictive performance changes with calibration set size and allows the identification of property-specific ranges beyond which additional calibration samples provide limited gains (Ng et al., 2020; Ramirez-Lopez et al., 2018). Crucially, this framework provides the empirical basis for optimizing sampling efficiency and evaluating data requirements under operational constraints.

Experiment 2 – Vertical Transfer (Topsoil to Subsoil)

To test whether calibrations developed on topsoil spectra can be applied to subsoil samples without depth-specific recalibration, models were trained exclusively on topsoil samples and validated on an independent set of subsoil samples. This vertical transfer design reflects a common practical scenario in which subsoil reference measurements are unavailable or scarce, and topsoil-calibrated models must be extended downwards. Topsoil and subsoil samples differ systematically in SOM content, mineral composition, and spectral character (Viscarra Rossel et al., 2022), making this an important test of model generalisability across depth. The degree to which

vertical transfer performance differs from depth-specific calibration provides evidence of whether topsoil-derived spectral models can be transferred to subsoil conditions or whether depth-specific calibration is required.

Experiment 3 – Temporal Transfer (2011 to 2021)

To assess the temporal transfer of spectral calibrations across a repeat soil survey, models were trained on samples from the 2011 dataset and validated on samples from the 2021 dataset, with no overlap between training and test sets. This design represents the operational scenario of applying a historical spectral calibration to a subsequent monitoring campaign without updating the model, thereby testing its temporal transferability under decadal changes in soil properties and spectral characteristics. Temporal transfer experiments of this type are increasingly recognized as a critical test for spectral monitoring frameworks intended for long-term soil change detection (Oldoni et al., 2024; Zayani et al., 2023).

Prediction Algorithms & Feature Selection

Prediction Algorithms

The regression algorithms were employed to develop spectral calibration models for the prediction of soil properties: partial least squares (PLS) regression, Cubist, and random forest (RF). These algorithms were selected to represent a range of modelling approaches, from linear latent-variable methods to rule-based and ensemble tree methods, ensuring a robust evaluation across varying degrees of complexity and non-linearity (Ng et al., 2019; Soriano-Disla et al., 2014). This multi-algorithmic framework provides the necessary methodological benchmark for scaling localized predictive structures up to more complex, cross-scale machine learning (ML) ensembles in subsequent landscape-level applications.

Partial least squares (PLS) regression decomposes the predictor matrix into a set of orthogonal latent variables that maximize the covariance between the spectral data and the response variable, thereby addressing the high multicollinearity inherent in soil spectra (Wold et al., 2001). PLS regression has long served as the standard benchmark in soil spectroscopic modelling owing to its robustness, parsimony, and interpretability (Viscarra Rossel & Behrens, 2010). The optimal number of latent variables was selected by minimising the root mean square error of cross-validation).

Cubist is a rule-based regression algorithm derived from Quinlan's M5 model tree (Quinlan, 1992), in which the predictor space is partitioned into a set of mutually exclusive rules and a separate multilinear regression model is fitted within each rule. Predictions are further refined by an instance-based correction mechanism that adjusts

311 model outputs using the most similar training cases (Quinlan, 1993). Cubist is particularly effective for soil spectral
312 data given its capacity to handle non-linear property–spectra relationships through localized multi-linear
313 approximation and has consistently performed well across a range of soil properties and spectral datasets (Hong
314 et al., 2022).

315 Random forest (RF) is an ensemble learning method that constructs a large number of decision trees from bootstrap
316 samples of the training data, with each tree split evaluated on a random subset of predictor variables (Breiman,
317 2001). Predictions are obtained by averaging across all trees, which reduces variance and improves generalization
318 relative to individual trees. RF is well-suited to high-dimensional spectral data and provides internal estimates of
319 variable importance via out-of-bag error, making it naturally amenable to feature selection (Breiman, 2001).

320 In the experiments explained above, within the calibration set, hyperparameter tuning was performed using k-fold
321 cross-validation (5-fold). For PLS regression, the number of latent variables (components) was tuned up to a
322 maximum of 15 to prevent overfitting, with the optimal number selected based on the lowest cross-validated
323 RMSE. For Cubist, the committee and neighbour parameters were tuned with a grid comprising committees (1,
324 10, 30, 50) and neighbours (0, 3, 5, 9), ranging from a single rule set without instance-based correction to 50
325 member committees with the nine neighbours (Song et al., 2026). For RF, the number of variables randomly
326 sampled at each split (mtry) was tuned across a grid ($\sqrt{p}$, p/5 and p/3) derived from the number of predictors (p;
327 spectral wavelengths), where p/3 correspond to default value for regression (Breiman, 2001). The number of trees
328 were set to default value of 500 as most studies reported error stabilization before reaching this value (Lawrence
329 et al., 2006). The train function in caret was used to tune and select the optimal parameters for each model and
330 dataset combination. The final model for each property was trained with hyperparameters based on the lowest
331 cross-validated RMSE and subsequently evaluated on the independent validation set.

332 *Feature selection algorithms*

333 To reduce spectral redundancy and improve model robustness, algorithm-specific feature selection methods were
334 applied to each prediction algorithm, alongside the full variable set (All Variables) as a baseline to assess whether
335 feature selection provided a similar or meaningful predictive advantage. Each algorithm was paired with a feature
336 selection method based on comparative evidence from soil Vis-NIR spectroscopy studies demonstrating that
337 specific feature selection methods perform better when matched with specific machine learning algorithms (Zhang
338 et al., 2023). Isolating these target-specific spectral features is essential to filter out environmental and
339 management-induced noise, directly enhancing model robustness against temporal and vertical domain drifts.

For PLS, the Competitive Adaptive Reweighted Sampling (CARS) algorithm was applied as the feature selection method. CARS iteratively selects wavelength subsets by simulating a Darwinian evolutionary process, retaining variables with the largest absolute regression coefficients from a PLS model while progressively eliminating uninformative ones (H. Li et al., 2009). Among the variable selection methods compared by (Zhang et al., 2023) across multiple predictive algorithms in a regional soil Vis-NIR spectral library, CARS consistently achieved the best performance when paired with PLS, supporting its selection as the paired feature selection method for PLS in the present study.

For Cubist, recursive feature elimination (RFE) was applied. RFE is a wrapper-based method that iteratively builds the model and removes the least important variables based on predictive performance, progressively converging on an optimal feature subset (Guyon & Elisseeff, 2003). RFE has been reported to achieve the best or near-best accuracy for Cubist-type models in soil spectroscopy applications (Zhang et al., 2023).

For RF, two feature selection methods were evaluated: Forward recursive feature selection (FRFS) and Boruta. Both methods exploit RF's internal variable importance measures, making them structurally consistent with the RF modelling framework. FRFS combines a forward selection strategy with recursive refinement, iteratively adding variables that improve model accuracy while progressively re-evaluating and eliminating less informative wavelengths, and has been identified as the best-performing feature selection method for RF (Zhang et al., 2023). Boruta is a wrapper algorithm built around RF that identifies all variables carrying statistically significant information relative to randomly permuted shadow features, providing a more conservative and comprehensive variable selection (Kursa & Rudnicki, 2010), and has similarly been shown to improve RF-based prediction accuracy for soil properties such as SOM ((Q. Zhou et al., 2023).

Analysis of selected features

To identify the most important spectral regions for each soil property, we extracted the wavelengths selected by the four feature selection methods (CARS, RFE, FRFS, and Boruta) for every soil property, depth, and spectral dataset. Each selected wavelength was assigned to either the Vis-NIR or MIR sensor and grouped into a spectral region according to known soil absorption features (Viscarra Rossel & Behrens, 2010). Spectral region boundaries and intervals were defined according to established soil spectroscopy literature to group fundamental molecular vibrations, electronic transitions, and characteristic absorption features (Soriano-Disla et al., 2014; Viscarra Rossel et al., 2006).

Feature selection methods differ widely in the wavelengths they select; comparing the raw number of selected bands favours those that retain more variables. To avoid this bias, we calculated the proportion of each method's selected bands falling within each spectral region and then averaged these proportions across the four methods. This gave a single selection proportion value for each spectral region, soil property, and depth, ensuring that every method contributed equally regardless of how many wavelengths it selected. Identifying these robust, invariant spectral diagnostic regions establishes the physiological and chemical rulesets required for transitioning from proximal ground-based sensors to upscale remote sensing diagnostics.

Model Evaluation

Model accuracy was evaluated using three complementary metrics: root mean square error (RMSE), the coefficient of determination (R^2), and the ratio of performance to deviation (RPD) using expressions 1-3. While RPD values of 1.4–2.0, 2.0–2.5, and >2.5 have been traditionally been interpreted as indicating acceptable, good, and excellent predictive performance, respectively (Chang et al., 2001), we acknowledge that rigid thresholds can be sensitive to the inherent variance of the target sample set. Thus, RPD is utilised here as a standardised metric for relative comparison across domains rather than an absolute classification of model utility.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(e_i - o_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(e_i - o_i)^2}{n}} \quad (2)$$

$$RPD = \frac{SD\ (obs)}{RMSE} \quad (3)$$

Where o_i and e_i represents observed and estimated values of the i^{th} sample respectively. $\bar{o}$ refers to the mean of the observed values, SD is standard deviation, and n is number of samples in analysis.

Statistical Analysis

Spatial Homogeneity & Temporal Stability Analysis

Spatial homogeneity and temporal stability of the experimental field were evaluated using spatial statistics, variance analyses, and multivariate spectral analyses for the years 2011 and 2021 using the R software environment (R Core Team, 2024).

Spatial Autocorrelation and Patterns

Spatial autocorrelation was assessed using Moran's I to determine whether soil properties exhibited structured spatial configurations. Moran's I values close to zero indicate a lack of significant spatial autocorrelation and therefore a spatially random distribution, whereas positive values indicate positive spatial autocorrelation and clustering. Statistical significance was evaluated using associated p-values, where non-significant results ($p > 0.05$) suggest that spatial patterns cannot be distinguished from randomness. Moran's I was additionally compared between topsoil and subsoil layers to evaluate differences in vertical structural continuity and to quantify the baseline structural boundary conditions before potential crop diversification trials.

Temporal Stability of Variance and Spatial Dependence

Changes in the overall magnitude of variance between 2011 and 2021 were analysed using Levene's test for equality of variances. To further quantify the magnitude and direction of temporal alterations in soil variability, the Coefficient of Variation (CV, %) and the Variance Ratio (VR; computed as the ratio of sample variance in 2021 to 2011) were calculated for each soil property and depth interval. Non-significant Levene-test results ($p > 0.05$) paired with a VR near 1.00 indicate stable absolute variance over time, suggesting that long-term agricultural management did not systematically alter the overall statistical heterogeneity of the field. Additionally, empirical kernel density distributions were compared to visually assess shape constancy and systematic temporal shifts in mean values. To evaluate whether long-term management altered the geostatistical structure and continuous organization of these properties over time, omnidirectional semi-variograms were modelled. Spatial dependence was evaluated using the Nugget-to-Sill ratio (Nugget/Sill), where ratios < 0.25 indicate strong spatial dependence (with a lower bound of 0.00 reflecting pure spatial continuity and a complete absence of unstructured random noise), 0.25 to 0.75 indicate moderate spatial dependence, and > 0.75 suggest a weak spatial structure dominated by stochastic random noise. Temporal comparison of these geostatistical configurations allowed us to determine whether anthropogenic operations induced a decoupling of soil properties from their initial spatial templates.

Spectral Stability

Principal component analysis (PCA) was applied to Vis-NIR and MIR spectral data to evaluate spectral stability over time. For each year, the mean Euclidean distance of samples to the PCA centroid was calculated as a measure of spectral dispersion. Similar centroid distances between years were interpreted as evidence of a stable variance structure despite potential temporal drift in absolute soil property values, thereby providing evidence for the temporal persistence of the core pedological fingerprint. To further quantify temporal stability at the level of the raw spectral signal, the mean spectrum of each year (2011 and 2021) was compared per depth and sensor (Vis-NIR and MIR) using two complementary metrics: the Spectral Angle Mapper (SAM) (Kruse et al., 1993),

which quantifies the spectral similarity between the two mean spectra independently of overall magnitude, and the Mean Absolute Difference (MAD) between corresponding bands. Smaller SAM angles and MAD values indicate closer agreement between years and, consequently, a high temporal stability of the spectral signal.

Statistical Comparison of Model Performance

To evaluate differences in soil property prediction performance across spectral datasets and machine-learning algorithms, we employed a non-parametric statistical framework. All analyses were stratified by soil depth (topsoil and subsoil) and performed using RPD as the evaluation metric. Non-parametric tests were chosen because the distribution of prediction performance metrics across diverse soil properties does not generally conform to normality.

Comparison of Spectral Datasets & Prediction Algorithms

We used the Friedman test, a non-parametric alternative to repeated-measures ANOVA, to identify significant differences in predictive performance. Separate one-factor tests were conducted: one comparing the spectral datasets (Vis-NIR, MIR, synergistic) and one comparing the algorithms (PLS, Cubist, RF), each performed separately for topsoil and subsoil. In this framework, the seven soil properties served as blocks, allowing us to account for inherent differences in spectral predictability among properties. Mean RPD values were used for the analysis. For the comparison of spectral datasets, RPD values were averaged across all algorithm and feature-selection combinations for each soil property and spectral dataset. For the comparison of algorithms, RPD values were averaged across all spectral datasets and feature-selection variants for each soil property and algorithm. Using these mean values characterises the typical performance of each spectral dataset or algorithm and prevents conclusions from being based on a single favourable combination.

Effect sizes were quantified using Kendall's coefficient of concordance (W) (Kendall, 1948), computed as:

$$W = \frac{{\chi_F}^2}{n\,(k-1)} \quad (4)$$

where ${\chi_F}^2$ is the Friedman test statistic, n is the number of blocks (seven soil properties), and k is the number of factor levels (three spectral dataset or three algorithms). W ranges from 0 to 1, with values interpreted as weak ($W < 0.30$), moderate ($0.30 \leq W < 0.50$), strong ($0.50 \leq W < 0.70$), and very strong ($W \geq 0.70$) concordance among rankings (Schmidt, 1997).

When the Friedman test indicated significant differences ($p < 0.05$), we conducted pairwise post-hoc comparisons using the Nemenyi test. This procedure effectively controls the familywise error rate and is the recommended

approach for ML benchmarks (Demšar, 2006). All analyses were performed using the PMCMRplus package (Pohlert, 2024) in R (R Core Team, 2024).

Feature-selection effects were evaluated separately for each prediction algorithm using paired Wilcoxon signed-rank tests, as the applied feature-selection methods were algorithm-specific and selected based on previous literature. For each algorithm, the predictive accuracy obtained with its corresponding feature-selection method (PLS–CARS, Cubist–RFE, RF–FRFS, and RF–Boruta) was compared with that obtained using the full pre-processed spectra. The Holm correction was applied across the family of feature-selection comparisons within each depth to control the familywise error rate. This paired design isolates the effect of reducing the spectral feature space while holding the prediction algorithm constant, allowing the contribution of feature selection to be evaluated independently of differences attributable to algorithm choice.

Results

This section presents the predictive performance of the spectral datasets, algorithms, and feature selection methods across both depths. It begins with the spatial homogeneity and temporal stability of the experimental field across topsoil and subsoil based on the soil properties analysis of data collected in 2011 and 2021. The three spectral datasets were then compared, followed by an evaluation of the ML algorithms and feature selection methods together with their interaction. Two further experiments follow: the sensitivity of model performance to calibration set size and the transferability of models across depths (vertical) and survey years (temporal). To establish a standardized baseline, the dataset and algorithm comparisons were both conducted using a fixed 70:30 train/test split, whereas the systematic variation of calibration set sizes (from 10:90 to 90:10) is evaluated separately in calibration size experiment (Experiment 1).

Spatial Homogeneity & Stability of the Crop Research Unit

Soil Property Distributions &Temporal Changes

The soil inventory data from 2011 and 2021 across the 15-ha Crop Research Unit Merbitz provide a detailed overview of the spatial distribution and decadal development of soil properties (**Table 1**).

The soil matrix across the site is predominantly silty, with silt fractions remaining stable around 79–80% in both topsoil (0–30 cm) and subsoil (30–60 cm). Minor decadal shifts include a slight increase in mean topsoil clay content (from 7.7% to 9.0%) and a corresponding slight decrease in sand fractions, while absolute variances for texture parameters remained largely constant.

Topsoil SOM concentrations showed a decrease from a mean of 4.22% in 2011 to 4.03% in 2021, coupled with an increase in spatial nugget variance. In contrast, subsoil SOM exhibited a minor yet significant increase from 3.28% to 3.51%.

EC exhibited a substantial overall decline across the profile, with mean values dropping by roughly 50% in both the topsoil (from 213.2 to 107.1 µS/cm) and the subsoil (from 216.4 to 130.4 µS/cm). Soil reaction showed a slight vertical convergence between layers: topsoil pH increased 7.2 to 7.4, whereas subsoil pH remained 7.7.

Spatial Autocorrelation & Field Structure

Moran's I analysis (**Table 2, Fig. 3**) revealed significant spatial autocorrelation for nearly all soil properties in both 2011 and 2021 ($p \approx 0$), indicating that the field was not perfectly homogeneous but characterized by persistent spatial gradients and clusters. Clay showed slightly weaker spatial autocorrelation in 2011 compared to the other parameters. Nevertheless, Moran's I values remained remarkably stable over the ten-year period, indicating temporal persistence in the spatial structure of the experimental field.

Physical properties such as clay, sand, silt, and SOM exhibited relatively low Moran's I values (approximately 0.10–0.20), indicating weak yet temporally persistent spatial structures across the two survey years. In contrast, chemical parameters such as pH showed stronger spatial autocorrelation (approximately 0.40–0.50), reflecting more pronounced gradients that were consistent between the survey years. Overall, both topsoil and subsoil maintained highly stable spatial patterns between 2011 and 2021 (**Table 2**, **Fig. 3**), indicating that the spatial organisation of the measured soil properties remained broadly consistent over the ten-year observation period.

Temporal Stability of Variance

Levene's test results (**Table 1**, **Fig. 4**) indicated no significant changes in variance for clay, sand, silt, and $CaCO_3$ in the topsoil ($p > 0.05$), indicating no detectable change in their overall variance between 2011 and 2021. Density distributions further showed highly consistent distribution shapes between 2011 and 2021 (**Fig. 4**). However, geostatistical indices revealed a shift in the spatial organization of SOM. While its variance remained statistically stable according to Levene's test, the coefficient of variation (CV) increased from 7.64% to 11.61%, and the Nugget/Sill ratio rose sharply from 0.00 to 0.62. Together, these results indicate an increase in the relative variability and a change in the spatial organisation of SOM between 2011 and 2021.

In contrast, pH and EC exhibited highly significant variance changes over time, accompanied by increased coefficients of variation in the topsoil (**Table 1**). For example, the CV for pH increased from approximately 5.5 to 8.8, indicating an increase in relative spatial variability between the two survey years. Interestingly, the Nugget/Sill

ratio for pH remained very low (0.04), suggesting that even though the magnitude of variability increased, its spatial organisation remained strongly structured.

In the subsoil, coefficients of variation for $CaCO_3$, clay, EC, sand, and silt decreased between 2011 and 2021 (**Table 1**), indicating a gradual homogenization of deeper soil layers. This trend is further supported by the Nugget/Sill ratios for silt and sand decreased from 0.20 and 0.16 in 2011 to 0.00 in 2021, indicating a decrease in the relative unstructured component of spatial variation and an increase in the relative contribution of spatially structured variation.

Spectral Stability

PCA-based spectral analyses (**Table 3**) demonstrated a highly stable spectral variance structure over the ten-year period. In the topsoil, the stability ratio of the mean distance to centroid was 1.01, indicating very similar overall spectral dispersion.

In contrast, the subsoil exhibited a subtle increase in spectral dispersion, yielding a stability ratio of 1.11 (**Table 3**). The PCA point cloud size expanded moderately (~11%) between 2011 and 2021. Rather than indicating structural instability, this indicates a moderate increase in overall spectral dispersion between the two survey years. This depth contrast was independently reflected in the mean spectra: both the spectral angle and the mean absolute difference between the 2011 and 2021 mean spectra were consistently smaller in the topsoil than the subsoil (Vis-NIR: SAM 2.24° vs. 3.90°, MAD 0.027 vs. 0.055; MIR: SAM 1.73° vs. 1.99°, MAD 0.020 vs. 0.027). These results indicate greater spectral similarity between survey years in the topsoil than in the subsoil across both spectral ranges.

Comparison of spectral datasets

The predictive performance of the three spectral datasets (Vis-NIR, MIR, and synergistic) was compared across all soil properties, models, and feature selection methods for both depths. **Fig. 5** and **Fig. 6** present the RPD performance based on a 70:30 train/test split for topsoil and subsoil, respectively, with corresponding R^2 and RMSE values provided in the Appendix (**Fig. A1**: Topsoil; **Fig. A2**: Subsoil). The Friedman test indicated a statistically significant effect of spectral dataset on predictive performance in both topsoil and subsoil (**Table 4**). The effect was larger in topsoil ($\chi^2 = 10.57$, df = 2, p = 0.005, Kendall's W = 0.76), compared with the subsoil ($\chi^2 = 7.14$, df = 2, p = 0.028, W = 0.51), indicating a lower degree of agreement in the relative performance of the spectral datasets across soil properties.

Post-hoc Nemenyi comparisons showed significant differences between the synergistic dataset and the single-range datasets at both depths (**Table 5**). In the topsoil, the synergistic dataset showed significantly higher predictive performance than MIR ($p = 0.009$) and Vis-NIR ($p = 0.021$), while the difference between Vis-NIR and MIR was not significant. In subsoil, the synergistic dataset showed significantly higher predictive performance than MIR ($p = 0.021$), whereas differences between synergistic and Vis-NIR ($p = 0.375$) and Vis-NIR and MIR ($p = 0.375$) were non-significant. These results performance difference associated with spectral data combination is more pronounced in the topsoil, where the combination of spectral ranges resulted in significantly higher predictive performance than either individual spectral range. In contrast, in the subsoil, the synergistic dataset showed significantly higher predictive performance only compared with MIR, while its performance did not differ significantly from Vis-NIR.

With respect to individual soil properties (encompassing SOM, texture fractions sand, silt, clay, $CaCO_3$, and pH_{H2O}), the synergistic dataset generally showed the highest or near-highest RPD values, particularly for carbonate and pH-related properties (**Fig. 5** and **Fig. 6**; comprehensive performance metrics for all individual models, datasets, and feature selection combinations are provided in Appendix A1 and A2). The highest RPD values were obtained for $CaCO_3$ (RPD up to 5.54 with Cubist), and pH_{H2O} (up to 3.07), with RPD values ≥ 2.0. Conversely, predictive performance for SOM and the texture fractions (clay, silt, sand) remained below RPD = 1.5 in the topsoil regardless of the dataset, with no substantial improvement associated with the use of the synergistic dataset. A notable exception occurred for the texture fractions in the subsoil, where silt and sand showed higher predictive performance (best RPD = 2.01 and 2.35, respectively) compared to the topsoil (1.35 and 1.32), while clay performance remained similar at both depths (RPD $\approx$ 1.3).

Overall, the results indicate that the relative performance of the spectral datasets differed between soil depths, with a more pronounced benefit of the synergistic dataset in the topsoil and higher RPD values for silt and sand in the subsoil.

Machine learning algorithms & feature selection

To evaluate the effect of ML algorithm choice, the Friedman test was conducted separately for each depth across all spectral datasets and soil properties. ML algorithm choice significantly influenced predictive performance in the subsoil, but not in the topsoil (Table 6a). In the subsoil, the effect showed a high effect size ($\chi^2 = 11.14$, df = 2, $p = 0.004$, Kendall's $W = 0.80$), whereas in the topsoil the differences among algorithms did not reach significance ($\chi^2 = 3.71$, df = 2, $p = 0.156$, $W = 0.27$). Post-hoc Nemenyi comparisons in the subsoil revealed that

Cubist showed significantly higher predictive performance than both RF ($p = 0.004$) and PLS ($p = 0.043$), while the difference between PLS and RF was not statistically significant (**Table 6**a). Cubist showed the highest mean RPD at both depths, followed by PLS and RF.

In contrast to the choice of algorithm, feature selection did not significantly affect predictive performance for any of the ML methods at either depth. Each feature-selection method was compared against the full pre-processed spectra within the same algorithm (CARS with PLS, RFE with Cubist, and FRFS and Boruta with RF) using the paired Wilcoxon signed-rank test with Holm correction (**Table 6**b); of the eight comparisons, only RF+Boruta in the subsoil reached significance, corresponding to a small difference in performance (mean RPD change ≤ 0.08). Algorithm-specific feature selection methods thus achieved predictive performance that was statistically similar to the full pre-processed spectra while utilizing only a reduced subset of wavelengths.

The magnitude and direction of the feature-selection effect varied across soil properties, spectral datasets, and depths (**Fig. 7**). Feature selection resulted in the largest positive changes in RPD for pH_{H2O}, with the largest gains for CARS applied to PLS (mean ΔRPD of +0.12 for pH_{H2O}), whereas effects on the textural fractions and SOM were small. For $CaCO_3$, which was already predicted relatively high RPD values from the full pre-processed spectra, feature selection tended to reduce performance, most markedly for RFE applied to Cubist in the subsoil. Thus, although feature selection does not yield a systematic its effects differed among soil properties, algorithms, and depths.

Spectral regions selected for individual soil properties

The number of variables retained differed markedly among methods (**Fig. 8**). CARS consistently selected the fewest wavelengths (~6–100), whereas RFE was the most variable, ranging from compact subsets to the full spectrum for SOM, clay, and silt. FRFS and Boruta retained intermediate numbers of variables with comparatively smaller variation among soil properties. The wavelengths selected by the four feature selection methods were grouped into spectral regions and summarized as selection proportions for each soil property and depth (**Fig. 9** and **Fig. 10**). The complete set of individual wavelengths selected by each method for every soil property and depth is provided in the Supplementary Material (**Fig.s S1: Vis-NIR; S2: MIR**), while **Fig. 9** and **Fig. 10** summarize these selections as region-level proportions.

The retention of the full feature set by RFE for SOM, clay, and silt indicates that RFE did not identify a smaller subset that reduced cross-validated error for these properties. In contrast, for the higher performing carbonate and

pH-related properties, all methods selected substantially reduced subsets, with fewer wavelengths retained than in the full pre-processed spectra.

$CaCO_3$ showed the most distinct selection pattern. In the MIR range it was the only property with notable selection in the carbonate region (1200–1600 cm^{-1}), with selection proportions of 0.21 in the topsoil and 0.28 in the subsoil. In the Vis-NIR range, $CaCO_3$ selection was concentrated in the visible and near-infrared-A regions (400–1000 nm), with proportions of 0.44 and 0.30, respectively, in the topsoil. The pH_{H2O} property showed high selection proportions in the Vis-NIR visible region (0.29–0.31) but lower selection proportions in the MIR carbonate region (0.12–0.13). Its highest MIR selection proportions were observed in the 2000–2800 cm^{-1} region (0.32–0.40).

For the texture fractions, clay selection was strongest in the Vis-NIR region (700–1000 nm; 0.34 in the subsoil) and the MIR silicate region (600–1200 cm^{-1}; 0.21 in the topsoil), with lower proportion in the 2100–2300 nm region (0.08–0.23). Silt and sand showed a broader selection across the visible and near-infrared regions. For silt and sand, selection patterns were more concentrated in the subsoil than in the topsoil, corresponding to the higher RPD values observed at depth.

In contrast, SOM exhibited a diffuse selection pattern with no single dominant region in either spectral range or depth, with selections distributed across the visible, near-infrared, and organic O–H regions.

Effect of calibration size

To evaluate how predictive performance changes with calibration sample size, the proportion of data allocated for model calibration was systematically varied from 10% to 90% (with the remaining fraction serving as the independent validation set). A standard 70:30 split was adopted for the primary evaluations, providing a fixed calibration and validation proportion for comparison across datasets, algorithms, and feature-selection methods. **Fig. 11** presents the results for each property using the RPD statistic; the corresponding R^2 and RMSE are provided in the appendix (**Fig. A3**). **Fig. 12** shows the highest observed performance, along with the corresponding model, feature selection method, and calibration proportion.

Predictive performance generally improved as the calibration proportion increased, but both the rate of improvement and the shape of the learning curve varied markedly among soil properties and between depths. In the topsoil, pH_{H2O}, $CaCO_3$ and EC stabilized at intermediate calibration proportions, beyond which additional samples produced only limited gains in RPD. In the subsoil, by contrast, the higher performing properties continued to improve across the full calibration range. Some curves showed declines at the highest calibration

proportions (e.g., 90%), most notably topsoil $CaCO_3$ and silt. At these calibration proportions, the validation set comprised only 10% of the observations, resulting in greater variability of the validation-based performance estimates.

While topsoil chemical properties showed early stabilisation, the texture fractions and SOM exhibited distinct learning trajectories. Unlike the chemical properties, topsoil texture fractions remained near the fair-prediction threshold across all training sizes, whereas subsoil sand and silt showed marked increases in RPD when calibration sets exceeded 70%. Because the validation set size changes reciprocally from 90% down to 10% alongside the training expansion, validation metrics at extreme subsets are inherently volatile. For instance, prediction of the sand fraction attained a RPD of up to 2.35 and silt up to 2.01 in the subsoil, compared with 1.32–1.35 in the topsoil. Thus, for these properties, larger calibration sets were associated with increasing RPD values at depth but not near the surface. $CaCO_3$ showed the highest RPD values, with an RPD of up to 5.54 in the subsoil for the MIR and synergistic datasets.

Overall, the chemical properties (pH_{H2O}, $CaCO_3$) in the topsoil achieved stable RPD values at moderate calibration sizes, whereas their subsoil predictions continued to improve across the full range. The more weakly predicted properties (SOM and the texture fractions in the topsoil) showed limited changes in RPD with additional calibration data. Across most properties, sensors, and depths, Cubist paired with the full variable set was among the highest-performing model configurations (**Fig. 12**), consistent with the ranking established earlier.

Transferability across time & depth

A key practical question for soil spectroscopy is whether a model calibrated under one set of conditions can be applied under another without new sampling and laboratory analyses. We evaluated this in two directions: vertical transfer, in which models calibrated using topsoil data were applied to subsoil data, and temporal transfer, in which models calibrated using the 2011 data were applied to the corresponding 2021 data. For both transfer scenarios, predictive performance was compared with the respective 70:30 baseline: the subsoil baseline for vertical transfer and the depth-specific baseline for temporal transfer. Full results for all model and feature-selection combinations are provided in the Appendix (**Fig. A4** and **Fig. A5**), while the highest observed performance for each soil property and spectral dataset is shown in **Fig. 13** and **Fig. 14** for vertical and temporal transfer, respectively. **Fig. 13**A and **Fig. 14**A present the RPD values for the highest-performing model for each soil property and spectral dataset, comparing the 70:30 baseline with the corresponding transfer performance. **Fig. 13**B and **Fig. 14**B show measured

versus predicted values for the overall highest-performing model for each property, with baseline and transfer predictions shown together.

For vertical transfer $CaCO_3$ showed the highest RPD values among the evaluated soil properties. Models calibrated on topsoil and applied to subsoil achieved RPD values above 3.0 for the MIR and synergistic datasets in both survey years (**Fig. 13**A). The corresponding scatter plots showed predicted values distributed closely around the 1:1 relationship, with regression lines comparable to those obtained for the respective 70:30 baseline models (**Fig. 13**B).

For the remaining soil properties, vertical transfer resulted in lower predictive performance than the corresponding subsoil baselines. For pH_{H2O}, RPD decreased from values above 2.0 in the 70:30 baseline to approximately 1.4–1.6 following transfer. EC, the texture fractions, and SOM showed further reductions, with RPD values below 1.4. The corresponding scatter plots showed regression lines with lower slopes than the 1:1 line, particularly for the transferred models, indicating a reduced spread of predicted values relative to the measured values. For pH_{H2O}, vertical transfer performance was slightly higher in 2021 than in 2011 and, to a lesser extent, the same pattern was observed for EC, resulting in higher transfer RPD values and smaller differences relative to the corresponding 70:30 baselines. No consistent difference between the two survey years was observed for the texture fractions, SOM, or $CaCO_3$.

Temporal transfer showed a broadly similar pattern. $CaCO_3$ again yielded the highest RPD values, exceeding 3.0 in the subsoil for the MIR and synergistic datasets (**Fig. 14**A). The remaining soil properties showed lower predictive performance under temporal transfer, with the magnitude of the reduction varying among properties. For EC, temporal transfer was associated with a pronounced shift between measured and predicted values, with predictions generally exceeding the corresponding measured values across the observed range (**Fig. 14**B). The texture fractions and pH_{H2O} showed patterns similar to those observed for vertical transfer, with predicted values concentrated towards the centre of the measured range. SOM consistently showed the lowest predictive performance, with RPD values of approximately 1.0.

Discussion

Temporal changes & stability of spatial structure over a decade

The results indicate that the experimental field is not perfectly homogeneous, but instead exhibits a persistent spatial organisation that remained broadly stable over the ten-year observation period. The persistence of significant spatial autocorrelation between 2011 and 2021 suggests that the field retained a coherent internal spatial

structure, consistent with previous observations of persistent soil spatial patterns in agricultural systems (e.g. Ahrends et al., 2023). This persistence indicates that the underlying spatial organisation was maintained despite a decade of agricultural management, including crop rotation, fertilisation, and tillage operations.

Importantly, the persistence of the spatial framework did not imply temporal stability of all soil properties. Instead, the geostatistical results revealed property-specific changes in the degree and organisation of spatial variability. This was particularly evident for SOM in the topsoil, for which the Nugget/Sill ratio increased from 0.00 in 2011 to 0.62 in 2021. This increase indicates a greater relative contribution of small-scale or unstructured variation in 2021 compared with 2011. The change may reflect the greater influence of temporally dynamic processes affecting topsoil organic matter, including crop-residue inputs, mineralisation and root-derived carbon inputs, which can modify the spatial distribution of SOM within a more persistent mineral soil framework. Such dynamics are consistent with the comparatively rapid turnover and spatial reorganisation of topsoil carbon reported by (Cotrufo & Lavallee, 2022).

A similar distinction between persistent spatial organisation and property-specific temporal change was evident for pH and EC. Although these properties showed changes in their distributions and relative variability, their spatial structures remained sufficiently organised to retain significant spatial autocorrelation. In particular, pH exhibited comparatively strong spatial autocorrelation in both survey years, suggesting that temporal changes in its magnitude did not eliminate its spatial organisation. This indicates that changes in soil properties can occur within a persistent spatial framework rather than necessarily replacing that framework. By contrast, the texture fractions showed comparatively limited temporal change in their spatial organisation, supporting the interpretation that properties associated more directly with the mineral soil matrix can exhibit greater spatial persistence than more dynamic chemical properties.

The spectral analyses provide a complementary perspective on this temporal stability. The PCA-based stability ratio was close to one for the topsoil (1.01), indicating very similar overall spectral dispersion between 2011 and 2021, whereas the higher ratio for the subsoil (1.11) indicated a moderate increase in spectral dispersion. The lower SAM and MAD values observed for the topsoil further indicated greater spectral similarity between survey years than in the subsoil. These results suggest that the temporal stability of the soil system was also reflected in its spectral properties, although the magnitude of spectral change differed with depth. Importantly, the spectral results should not be interpreted as evidence that agricultural management completely homogenised the topsoil or that the subsoil remained unaffected. Rather, they indicate that substantial changes in individual soil properties occurred within a spatial and spectral framework that retained considerable temporal continuity.

The comparatively persistent spatial and spectral characteristics observed in the subsoil suggest that deeper soil layers may provide a comparatively stable reference context for long-term soil monitoring. However, the observed temporal changes indicate that the subsoil should not be regarded as an unaffected or static archive.

Taken together, these findings provide support for the first hypothesis, which proposed that relatively stable pedological properties would retain their spatial structure over the decadal observation period, while more dynamic properties would exhibit greater temporal variability. They also provide partial support for the expected depth-dependent differences in the second hypothesis. The topsoil showed clear changes in the spatial variability of dynamic properties such as SOM, while the subsoil exhibited its own changes, including shifts in relative variability and spatial structure. Thus, the results do not support a simple distinction between a homogenised topsoil and an unchanged subsoil. Instead, they point to depth-dependent modes of temporal change superimposed on a persistent spatial framework. This distinction is important for soil spectroscopy because temporal stability of the spatial framework does not necessarily imply stability of the property-spectra relationships, providing a basis for examining sensor performance and model transferability in the following sections.

Spectral modelling & performance

The comparison of the Vis-NIR, MIR and synergistic datasets demonstrated that spectral dataset selection significantly affected predictive performance, although the magnitude of this effect depended on soil depth and target property. In the topsoil, the synergistic dataset showed significantly higher predictive performance than both MIR and Vis-NIR individually, whereas Vis-NIR and MIR did not differ significantly. In the subsoil, the synergistic dataset also performed significantly better than MIR but did not differ significantly from Vis-NIR. This depth-dependent pattern suggests that combining spectral ranges can provide complementary information, but that the benefit of combining spectral ranges is not uniform across soil layers or target properties. The results are consistent with previous studies reporting improved soil-property prediction when complementary spectral information is combined (Nocita et al., 2015; Viscarra Rossel et al., 2006), while also demonstrating that the magnitude of the benefit depends on the spectral characteristics and variability of the target property.

The strongest benefit of the synergistic dataset was observed for $CaCO_3$ and, to a lesser extent, pH_{H2O}. $CaCO_3$ achieved the highest predictive performance among the investigated properties, with RPD values of up to 5.54. This strong performance is consistent with the distinct fundamental absorption features of carbonate minerals in the MIR region, which provide a more direct spectral basis for their quantification than is available for many other soil properties (Comstock et al., 2019; Mammadov et al., 2024). The frequent selection of wavelengths in the

carbonate-related spectral regions further supports the importance of these regions for the $CaCO_3$ models. In particular, the selected MIR ranges around 1200-1600 cm^{-1} and in the 2000–2800 cm^{-1} encompass characteristic carbonate absorption features, including the asymmetric C–O stretching region and carbonate combination bands (Comstock et al., 2019). The combination with Vis-NIR therefore appears to provide complementary spectral information rather than simply increasing the number of available wavelengths.

The comparatively good prediction of pH_{H2O} should be interpreted differently. Soil pH does not possess a strong, unique absorption feature that can be directly related to its concentration in the same way as $CaCO_3$. Its prediction is therefore likely to depend on covariance with spectrally active soil constituents, including carbonates and mineralogical components, as reported in digital soil mapping studies (de Santana et al., 2025; Greenberg et al., 2023). The observed selection of spectral regions associated with these constituents is consistent with such indirect relationships. However, because the present study did not experimentally separate these covariance pathways, the selected wavelengths should be interpreted as indicators of predictive relevance rather than as direct evidence of a specific causal spectral mechanism.

In contrast, SOM, EC and the texture fractions remained more difficult to predict, particularly in the topsoil. For SOM, the comparatively narrow concentration range and low coefficient of variation within the Merbitz experimental field likely limited the spectral contrast available for calibration. This interpretation is consistent with previous studies showing that spectroscopic prediction performance depends strongly on the range and diversity of the calibration dataset (Ng et al., 2020; Oberholzer et al., 2024). Thus, low predictive performance for SOM should not necessarily be interpreted as a lack of spectral sensitivity, but rather as a consequence of the restricted variability represented in the present field-scale dataset. For SOM, the absence of a consistently concentrated spectral region coincided with its comparatively low predictive performance and, in some cases, the retention of a large proportion of the available features by RFE. Together, these results suggest that the predictive information for SOM in the present dataset was relatively weak and distributed across a broad spectral range rather than concentrated in a small number of clearly identifiable regions.

The comparatively weak prediction of clay, silt and sand fractions can likewise be related to the indirect nature of their spectral estimation. Diffuse-reflectance spectroscopy does not measure particle size directly; instead, predictions can exploit covariance between particle-size fractions and spectrally active constituents such as clay minerals, organic matter and iron-bearing minerals (Janik et al., 1998; Tümsavaş et al., 2019). The stronger performance of subsoil sand and silt compared with their topsoil counterparts may therefore reflect differences in

766 the composition and variability of the respective datasets. However, this result should be interpreted as a site- and
767 dataset-specific observation rather than evidence that subsoil texture is generally more amenable to spectroscopic
768 prediction.

769 The model comparison further showed that Cubist achieved the highest mean predictive performance across the
770 evaluated configurations and significantly outperformed RF and PLS in the subsoil. This may reflect the ability of
771 Cubist to represent nonlinear relationships while retaining local linear approximations within individual rules.
772 Such flexibility can be advantageous for soil-spectral relationships that vary across the predictor space. The
773 observed ranking is consistent with previous studies reporting competitive or improved performance of Cubist for
774 soil-property prediction compared with conventional linear approaches and tree-based models (Viscarra Rossel et
775 al., 2006; Zhang et al., 2023). However, because the algorithm comparison was based on predictive performance
776 rather than an analysis of the underlying model structure, the present results do not allow the relative contribution
777 of specific algorithmic characteristics to be isolated.

778 Feature selection produced statistically similar predictive performance to the full spectral datasets across almost
779 all evaluated model combinations, despite substantially reducing the number of wavelengths used. This indicates
780 that a considerable proportion of the available spectral variables was redundant for the prediction tasks considered
781 here. Feature selection therefore provided more compact modelling inputs without a corresponding loss in
782 predictive performance. This reduction in spectral dimensionality may also facilitate data processing in
783 applications involving large spectral datasets, although such computational benefits were not evaluated directly in
784 the present study. The result is consistent with previous findings that spectral dimensionality can be reduced
785 without necessarily compromising predictive performance (Zhang et al., 2023). The selected spectral regions also
786 differed among soil properties, with $CaCO_3$ showing more concentrated selection around carbonate-related MIR
787 regions, whereas SOM exhibited a broader distribution of selected wavelengths. This property-specific selection
788 pattern supports the interpretation that feature selection can identify compact subsets of informative spectral
789 variables, although the selected wavelengths should not automatically be interpreted as unique causal absorption
790 features.

791 Overall, the modelling results demonstrate that spectral prediction performance was governed by an interaction
792 between sensor configuration, soil depth, target-property variability and modelling approach. Combination of Vis-
793 NIR and MIR information provided measurable benefits, particularly for properties with strong or complementary
794 spectral information, while properties with limited variability or predominantly indirect spectral relationships

remained more challenging. This property-dependent behaviour is important for interpreting the subsequent transferability analyses, because temporal or vertical model transfer can be expected to depend not only on the stability of the soil properties themselves but also on the stability of the spectral relationships used to predict them.

Sample size & performance

Increasing the proportion of samples allocated to calibration generally improved predictive performance, but the magnitude and shape of this response varied substantially among soil properties and depths. This property-dependent behaviour is consistent with previous studies showing that the amount of calibration data required for soil spectroscopy depends on the target property and the characteristics of the calibration dataset (van Leeuwen et al., 2026). In the present study, properties with comparatively strong and distinct spectral relationships, particularly $CaCO_3$ and pH_{H2O}, reached relatively high RPD values at small to moderate calibration proportions. Subsoil $CaCO_3$ exceeded an RPD of 3.0 even at a calibration proportion of 10%, whereas topsoil $CaCO_3$ reached this level at approximately 20-30%. In contrast, subsoil pH_{H2O} and sand required substantially larger calibration proportions, with performance increasing more clearly at approximately 60-70% and above 70%, respectively. These differences indicate that calibration requirements are strongly property-dependent rather than determined by sample size alone.

For several well-predicted properties, predictive performance approached a plateau at intermediate calibration proportions, indicating that additional calibration samples provided only limited gains beyond a certain point. This was particularly evident for topsoil pH_{H2O}, $CaCO_3$ and EC. Such behaviour suggests that once the main range of relevant spectral and property variability is adequately represented, further increasing the calibration set may provide diminishing improvements in predictive performance. However, this should not be interpreted as evidence that a specific calibration proportion is generally sufficient, because the observed threshold depends on the property, depth and underlying variability of the dataset. (Y. Zhou et al., 2026) similarly reported that the optimal calibration size was influenced by the soil property, the calibration algorithm and the spatial scale. Moreover, the calibration proportions evaluated here refer to the fraction of samples assigned to model calibration and should therefore not be equated directly with spatial sampling density or the number of field samples required for operational surveys.

The contrast between chemical properties and texture fractions further illustrates the importance of the underlying predictability of the target variable. SOM and the topsoil texture fractions remained close to or below the threshold for useful prediction across the investigated calibration proportions. Increasing the calibration dataset therefore

did not substantially improve their performance. This suggests that additional samples alone cannot compensate for a restricted spectral signal or limited property variability. For these properties, improvements in calibration performance may instead require a dataset that encompasses a broader and more informative range of soil conditions. In contrast, the marked increase in subsoil sand and silt prediction at larger calibration proportions indicates that additional calibration data can be beneficial where sufficient spectral information is present, but a larger sample is required to represent the relevant variability. This is consistent with the observation by (Ng et al., 2018) that the optimum calibration sample size depends on the degree of generalization the model must achieve.

The decline in performance observed for some properties at the highest calibration proportion should also be interpreted cautiously. At a calibration proportion of 90%, only 10% of the observations remained available for independent validation, increasing the sensitivity of performance estimates to the composition of the validation subset. This is particularly relevant for properties such as topsoil $CaCO_3$ and silt, for which RPD decreased at the highest calibration proportion despite generally increasing performance at lower proportions. The result therefore does not necessarily indicate that adding calibration samples reduced model quality; rather, it highlights the trade-off between calibration size and the amount of independent data available for evaluating model performance.

From a methodological perspective, the results support a shift from considering calibration size as a simple question of sample quantity towards considering the representativeness of the calibration dataset. A relatively small calibration set can be sufficient for properties with strong and well-defined spectral relationships, whereas properties with weaker or more variable relationships may require substantially more observations. Previous work has similarly emphasised the importance of selecting calibration samples that adequately represent the relevant covariate space rather than relying solely on large sample numbers (Guerrero et al., 2010). In the present study, however, the calibration subsets were generated through stratified random sampling and the analysis did not explicitly test alternative spatial or covariate-optimised sampling strategies. Consequently, the results provide evidence for property-dependent calibration requirements, but do not directly establish an optimal spatial sampling design.

Overall, the calibration-size analysis complements the modelling results by showing that predictive performance depends on both the information contained in the spectral data and the amount of calibration information used to learn the respective relationships. For strongly predictable properties such as $CaCO_3$, relatively small calibration sets captured much of the achievable predictive performance, whereas increasing calibration size provided more limited benefits for properties constrained by weak spectral relationships or restricted variability. This distinction

853 is relevant for the development of efficient soil-monitoring strategies, where reducing unnecessary calibration
854 samples could lower analytical effort while maintaining predictive performance, provided that the selected samples
855 adequately represent the relevant range of soil and spectral variability.

856 **Modeling transferability**

857 The transferability analyses provide an important link between the temporal and spatial stability of the soil system
858 and the stability of the corresponding spectral prediction models. Across both vertical and temporal transfer,
859 predictive performance was strongly property-dependent. Properties that were well predicted under the respective
860 70:30 baseline generally retained higher performance after transfer, whereas properties with weaker baseline
861 performance showed larger losses. $CaCO_3$ represented the clearest example of this pattern, maintaining
862 comparatively high RPD values under both vertical and temporal transfer, particularly for the MIR and synergistic
863 datasets. In contrast, SOM showed consistently low transfer performance, with RPD values close to 1.0. This
864 relationship suggests that transferability was closely linked to the strength and stability of the property-spectra
865 relationship rather than being determined by temporal or vertical distance alone.

866 The particularly strong transferability of $CaCO_3$ is consistent with the distinct fundamental absorption features of
867 carbonate minerals. Because these features provide a comparatively direct spectral basis for prediction, their
868 representation in the calibration data appears to remain sufficiently consistent across the investigated depths and
869 survey years. Similar transfer behaviour has been reported for geogenic soil properties with pronounced spectral
870 features, including inorganic carbon (Sanderman et al., 2023). In contrast, properties such as pHH2O and the
871 texture fractions are estimated more indirectly through covariance with spectrally active soil constituents. Such
872 relationships can be more sensitive to changes in soil composition and the distribution of covarying constituents,
873 potentially reducing transfer performance when the calibration and application datasets differ (Broeg et al., 2023).
874 The present results therefore support the interpretation that transferability depends not only on the temporal or
875 spatial stability of the target property itself, but also on the stability of the spectral relationships used to predict it.

876 Vertical transfer generally showed greater reductions in predictive performance than temporal transfer in the
877 present dataset, although the magnitude of this difference varied among properties. This pattern was particularly
878 apparent for pH_{H2O}, for which models calibrated on topsoil data showed a marked reduction in performance when
879 applied to subsoil data. The stronger challenge associated with vertical transfer is consistent with the substantial
880 differences in soil composition and structure between topsoil and subsoil, which can alter the covariance structure
881 underlying spectral prediction. These differences may include contrasting organic matter contents, mineralogical

882 composition and aggregation, as well as changes in the relative contribution of spectrally active constituents. The
883 observed pattern is therefore consistent with the interpretation that changes across soil depth can alter the property–
884 spectra relationship more strongly than temporal changes within the same depth over the ten-year observation
885 period (Renuka et al., 2026). At the same time, the broadly persistent spatial structure observed between 2011 and
886 2021 provides a plausible context for the comparatively higher temporal transferability at a given depth.

887 The transfer predictions also revealed systematic differences in the distribution of predicted values. Regression
888 relationships for several transferred models showed flatter slopes than the 1:1 relationship, indicating a reduced
889 spread of predicted values relative to the measured values. This pattern is consistent with the tendency of models
890 calibrated on one distribution to compress predictions when applied to data with a different range or covariance
891 structure. EC provided the clearest example of directional temporal bias, with predictions generally exceeding the
892 corresponding measured values in 2021. The mean EC values decreased substantially between the two survey
893 years, indicating a temporal shift in the target-property distribution that may have contributed to the observed
894 transfer bias. However, because the present analysis does not isolate the individual causes of this shift, the
895 relationship should be interpreted as an association rather than as direct evidence that the change in mean EC alone
896 caused the transfer error.

897 Interestingly, some transferred models performed slightly better in 2021 than in 2011. This was observed
898 particularly for pH_{H2O} and, to a lesser extent, EC in vertical transfer, and coincided with a reduction in relative
899 variability for some of these properties. Such patterns indicate that transfer performance cannot be predicted solely
900 from the direction or magnitude of temporal change in a soil property. Instead, the compatibility between the
901 calibration distribution and the target distribution, including their variance and covariance structures, appears to
902 be important. The absence of a consistent difference between the two survey years for texture, SOM and $CaCO_3$
903 further supports the property-specific nature of transferability.

904 Taken together, these findings provide partial support for the third hypothesis. The persistence of spatial and
905 spectral characteristics across the ten-year observation period was sufficient to support transfer for some properties,
906 particularly $CaCO_3$, but did not guarantee successful transfer for all target variables. Instead, transferability
907 depended on how stable the relationship between the spectral predictors and the target property remained across
908 depth and time. Properties with strong and comparatively direct spectral signatures were less sensitive to domain
909 changes, whereas properties characterised by weaker or more indirect relationships were more susceptible to shifts
910 in the calibration-application relationship. This distinction highlights an important limitation of assuming that a

911 stable field-scale soil structure automatically results in transferable spectral models: structural persistence can
912 provide a favourable basis for transfer, but the stability of the property-spectra relationship ultimately determines
913 whether this potential is realised.

914 **Implications and limitations**

915 The findings of this study have several implications for the development of efficient soil-spectroscopy-based
916 monitoring strategies. Combining Vis-NIR and MIR information improved predictive performance for several soil
917 properties, with particularly pronounced benefits observed for $CaCO_3$ and, in some cases, pH_{H2O}. However, the
918 magnitude of the benefit was property- and depth-dependent. This suggests that combination of Vis-NIR and MIR
919 information should not be regarded as universally advantageous, but rather as a strategy whose value depends on
920 the target properties and the information contained in the individual spectral domains. For properties with strong
921 spectral features, such as $CaCO_3$, MIR provided substantial predictive information on its own, whereas combining
922 spectral ranges provided additional benefits for some properties and depths.

923 This distinction is also relevant for long-term spectroscopic modelling, because persistent spatial structure alone
924 does not ensure that the relationship between spectral predictors and soil properties remains stable over time. Long-
925 term applications therefore require explicit evaluation of temporal stability and transferability rather than assuming
926 that a calibration developed under one set of soil conditions will remain unchanged.

927 The calibration-size analysis further indicates that efficient model development depends on the representativeness
928 of the calibration dataset rather than on sample number alone. Several properties reached comparatively high
929 predictive performance at relatively small calibration proportions, while others showed limited improvement
930 despite increasing the calibration set. This distinction has practical implications for reducing unnecessary
931 laboratory analyses during model development. However, the present study did not explicitly compare alternative
932 sampling designs or optimise the spatial distribution of calibration samples. The calibration subsets were generated
933 using stratified random sampling, and the effect of spatially explicit or covariate-based sampling strategies
934 therefore remains an important subject for future research. In particular, future studies could evaluate whether
935 samples selected to represent the relevant spectral and soil-property space provide equivalent or improved
936 predictive performance with fewer observations (Dorantes et al., 2022).

937 The comparatively strong temporal and vertical transferability of $CaCO_3$ indicates that models associated with
938 relatively stable spectral relationships can retain predictive performance across the investigated survey years and

soil depths. In contrast, the reduced transferability of SOM, pH_{H2O} and several texture fractions indicates that a single calibration dataset may not be sufficient when the covariance structure between spectral variables and target properties changes. For such properties, periodic validation and, where necessary, additional calibration samples may be required to maintain predictive performance. This suggests that future monitoring frameworks should distinguish between properties that can be monitored using relatively stable calibration relationships and those that require more frequent evaluation or recalibration.

An important limitation of the study is that the analysis was conducted at a single long-term experimental site with comparatively uniform parent material and a defined management history. This setting provides an important advantage for investigating temporal changes because persistent site characteristics and experimental management can be examined under relatively controlled conditions. At the same time, the limited environmental range restricts the extent to which the resulting calibration and transfer relationships can be generalised to soils with different parent materials, textures, climatic conditions or management histories. The observed transferability therefore represents transfer within the specific environmental and pedological context of the Merbitz site rather than a demonstration of regional or continental model transfer.

Despite these limitations, the combination of repeated soil sampling, spatial analysis and proximal spectroscopy provides a useful framework for linking changes in soil properties with the temporal stability of their spectral representation. The results suggest that future monitoring approaches could benefit from integrating proximal spectroscopy with spatially extensive remote-sensing observations. Satellite missions such as Sentinel-2 and spaceborne imaging spectroscopy platforms such as EnMAP and PRISMA offer opportunities to extend field-based information across larger spatial domains (Mahmood et al., 2026; Ward et al., 2024). However, such upscaling will require additional validation because the relationships identified at the field scale may not remain unchanged across broader environmental gradients. A promising pathway is therefore a hierarchical monitoring strategy in which targeted field-based sampling and laboratory spectroscopy are used to maintain and validate calibration relationships, while remote sensing provides spatially extensive observations between field campaigns.

Overall, the results support the use of soil spectroscopy as a component of long-term soil monitoring, while highlighting that its effectiveness depends on the stability of both the target properties and the associated spectral relationships. The combination of Vis-NIR and MIR information, property-specific calibration strategies and transferability assessment provides a basis for reducing unnecessary sampling while retaining independent validation of model performance. Rather than assuming that one calibration framework can be transferred

unchanged across all properties, depths and time periods, future monitoring systems should account explicitly for differences in spectral behaviour, temporal variability and transferability among soil properties.

Conclusion

This study investigated the temporal and spatial stability of soil properties and their implications for Vis-NIR, MIR and synergistic soil spectroscopy at the long-term experimental field site Merbitz in Central Germany. By combining geostatistical analysis, spectral comparison, machine-learning modelling, calibration-size analysis and model transfer across soil depths and survey years, the study assessed how persistent soil structure and temporal property changes affect spectroscopic prediction and transferability.

The results showed that the spatial framework of the field persisted over the ten-year observation period, while the magnitude and organisation of individual soil properties changed to different degrees. Relatively stable properties retained more persistent spatial patterns, whereas more dynamic properties, particularly topsoil SOM, showed stronger changes in their spatial variability. The results therefore support the hypothesis that temporal change occurs within a persistent spatial framework rather than replacing it. They also indicate that depth-related differences in temporal behaviour are more nuanced than a simple distinction between a homogenised topsoil and an unchanged subsoil.

Spectral modelling further showed that the benefit of combining Vis-NIR and MIR information was property- and depth-dependent. The synergistic dataset achieved significantly higher predictive performance than MIR in both depths and additionally outperformed Vis-NIR in the topsoil. $CaCO_3$ was the most accurately predicted property, with RPD values of up to 5.54, whereas SOM and the topsoil texture fractions remained comparatively difficult to predict. Cubist achieved the highest mean predictive performance among the evaluated algorithms, while feature selection generally maintained comparable performance with substantially fewer wavelengths. These findings demonstrate that sensor configuration, target-property variability and model choice jointly determine spectroscopic performance.

The calibration-size analysis showed that the amount of calibration data required was strongly dependent on the target property. Increasing the calibration proportion improved predictive performance for several properties, but gains became limited once the relevant variability was sufficiently represented. In contrast, increasing calibration size provided little improvement for properties with weak predictive performance, particularly SOM and topsoil texture fractions. Thus, increasing the number of calibration samples cannot fully compensate for limited spectral

information or restricted property variability, highlighting the importance of representative calibration datasets in addition to sample number.

Model transferability was likewise strongly property-dependent. Properties with comparatively stable spectral relationships, most clearly $CaCO_3$, retained higher predictive performance across soil depths and survey years, whereas properties relying on weaker or more indirect spectral relationships were more sensitive to changes in depth and time. The results therefore provide partial support for the hypothesis that persistence of the soil system can facilitate temporal and vertical model transfer, but also demonstrate that structural or spectral stability alone does not guarantee transferability. The stability of the property-spectra relationship is a critical additional requirement.

Overall, the findings demonstrate that long-term soil monitoring with proximal spectroscopy cannot be based on a single, universally transferable calibration strategy. Instead, monitoring approaches should account for differences among soil properties in spatial variability, spectral predictability, calibration requirements and transferability. The combination of multi-sensor spectroscopy, property-specific calibration strategies and explicit assessment of temporal and vertical transferability provides a basis for developing more efficient long-term soil monitoring approaches while retaining validation of model performance.

Acknowledgements

Many thanks to B. Look, M. Herfurth, S. Booth, R. Gensler and M. Ahrens of the Crop Research Unit Merbitz for their attempt to implement the trial and general help. We gratefully acknowledge the expert technical assistance and guidance of M. von Hoff and S. Kurze and are grateful to R. Backhaus, D. Teubner, S. Walch and S. Lange for their support in the field and laboratory work.

Conflict of interest

Authors declare that they have no known financial or personal interest that could influence the work reported in this paper.

Funding Declaration

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Ahrends, H. E., Simojoki, A., & Lajunen, A. (2023). Spatial pattern consistency and repeatability of proximal soil sensor data for digital soil mapping. *European Journal of Soil Science*, *74*(5), e13409. https://doi.org/10.1111/ejss.13409

Amundson, R., & Jenny, H. (1997). On a state factor model of ecosystems. *BioScience*, *47*(9), 536–543. https://doi.org/10.2307/1313122

Baldacci, A., Malathy Devi, V., Chen, D.-W., Narahari Rao, K., & Fridovich, B. (1978). Absorption spectrum of carbon dioxide at 4.3 μm. *Journal of Molecular Spectroscopy*, *70*(1), 143–159. https://doi.org/10.1016/0022-2852(78)90016-4

Barra, I., Haefele, S. M., Sakrabani, R., & Kebede, F. (2021). Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. *TrAC Trends in Analytical Chemistry*, *135*, 116166. https://doi.org/10.1016/j.trac.2020.116166

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Broeg, T., Blaschek, M., Seitz, S., Taghizadeh-Mehrjardi, R., Zepp, S., & Scholten, T. (2023). Transferability of covariates to predict soil organic carbon in cropland soils. *Remote Sensing*, *15*(4), 876. https://doi.org/10.3390/rs15040876

Cambardella, C. A., & Karlen, D. L. (1999). Spatial analysis of soil fertility parameters. *Precision Agriculture*, *1*(1), 5–14. https://doi.org/10.1023/A:1009925919134

Chang, C.-W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. *Soil Science Society of America Journal*, *65*(2), 480–490. https://doi.org/10.2136/sssaj2001.652480x

Comstock, J. P., Sherpa, S. R., Ferguson, R., Bailey, S., Beem-Miller, J. P., Lin, F., et al. (2019). Carbonate determination in soils by mid-IR spectroscopy with regional and continental scale models. *PloS One*, *14*(2), e0210235. https://doi.org/10.1371/journal.pone.0210235

Cotrufo, M. F., & Lavallee, J. M. (2022). Chapter One - Soil organic matter formation, persistence, and functioning: A synthesis of current understanding to inform its conservation and regeneration. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 172, pp. 1–66). Academic Press. https://doi.org/10.1016/bs.agron.2021.11.002

de Santana, F. B., Bondi, G., & Daly, K. (2025). Pathways to improving MIR prediction accuracy for soil carbon, nitrogen and pH using external libraries at national scale. *Geoderma*, *460*, 117410. https://doi.org/10.1016/j.geoderma.2025.117410

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

DIN e.V. (1994). *Soil quality – Determination of the specific electrical conductivity (ISO 11265:1994 + ISO 11265:1994/Corr.1:1996)*. Deutsches Institut für Normung e.V. https://doi.org/10.31030/7330547

DIN e.V. (1995). *Soil quality – Determination of carbonate content – Volumetric method (ISO 10693)*. Deutsches Institut für Normung e.V. https://doi.org/10.31030/2143356

DIN e.V. (2021). *Soil, treated biowaste and sludge – Determination of pH (ISO 10390)*. Deutsches Institut für Normung e.V. https://doi.org/10.31030/3226922

Dorantes, M. J., Fuentes, B. A., & Miller, D. M. (2022). Calibration set optimization and library transfer for soil carbon estimation using soil spectroscopy—A review. *Soil Science Society of America Journal*, *86*(4), 879–903. https://doi.org/10.1002/saj2.20435

FAO. (1998). *World Reference Base for soil resources* (No. 84; World Soil Resources Report). ISSS–ISRIC–FAO.

FAO. (2024). *A primer on soil analysis using visible and near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy*. FAO, Rome. https://doi.org/10.4060/cb9005en

Greenberg, I., Vohland, M., Seidel, M., Hutengs, C., Bezard, R., & Ludwig, B. (2023). Evaluation of mid-infrared and X-ray fluorescence data fusion approaches for prediction of soil properties at the field scale. *Sensors*, *23*(2), 662. https://doi.org/10.3390/s23020662

Guerrero, C., Zornoza, R., Gómez, I., & Mataix-Beneyto, J. (2010). Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma*, *158*(1), 66–77. https://doi.org/10.1016/j.geoderma.2009.12.021

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182. https://doi.org/10.1162/153244303322753616

Hong, Y., Munnaf, M. A., Guerrero, A., Chen, S., Liu, Y., Shi, Z., et al. (2022). Fusion of visible-to-near-infrared and mid-infrared spectroscopy to estimate soil organic carbon. *Soil and Tillage Research*, *217*. https://doi.org/10.1016/j.still.2021.105284

Janik, L. J., Merry, R. H., & Skjemstad, J. O. (1998). Can mid infrared diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture*, *38*(7), 681–696. https://doi.org/10.1071/EA97144

Jenny, H. (1941). *Factors of soil formation: A system of quantitative pedology*. McGraw-Hill.

Kendall, M. G. (1948). *Rank correlation methods* (1st ed.). Charles Griffin, London, UK.

Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heidebrecht, K. B., Shapiro, A. T., Barloon, P. J., et al. (1993). The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, *44*(2), 145–163. https://doi.org/10.1016/0034-4257(93)90013-N

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Kuzyakov, Y., & Zamanian, K. (2019). Reviews and syntheses: Agropedogenesis – Humankind as the sixth soil-forming factor and attractors of agricultural soil degradation. *Biogeosciences*, *16*(24), 4783–4803. https://doi.org/10.5194/bg-16-4783-2019

Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, *100*(3), 356–362. https://doi.org/10.1016/j.rse.2005.10.014

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, *648*(1), 77–84. https://doi.org/10.1016/j.aca.2009.06.046

Li, S., Shen, X., Shen, X., Cheng, J., Xu, D., Makar, R. S., et al. (2025). Improving the accuracy of soil classification by using vis–NIR, MIR, and their spectra fusion. *Remote Sensing*, *17*(9). https://doi.org/10.3390/rs17091524

Mahmood, T., Conrad, C., Wenzel, J. L., Ward, K. J., Chabrillat, S., & Pöhlitz, J. (2026). Improving soil organic carbon estimation by combining multi-temporal optical and SAR data. *Geoderma*, *468*, 117787. https://doi.org/10.1016/j.geoderma.2026.117787

Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, *154*(1–2), 138–152. https://doi.org/10.1016/j.geoderma.2009.10.007

Mammadov, E., Denk, M., Mamedov, A. I., & Glaesser, C. (2024). Predicting soil properties for agricultural land in the Caucasus Mountains using mid-infrared spectroscopy. *Land*, *13*(2), 154. https://doi.org/10.3390/land13020154

Minasny, B., & McBratney, Alex. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., & Mouazen, A. (2016). Estimating the soil clay content and organic matter by means of different calibration methods of Vis-NIR diffuse reflectance spectroscopy. *Soil and Tillage Research*, *155*, 510–522.

Ng, W., Minasny, B., Malone, B., & Filippi, P. (2018). In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra. *PeerJ*, *6*, e5722. https://doi.org/10.7717/peerj.5722

Ng, W., Minasny, B., Mendes, W. de S., & Demattê, J. A. M. (2020). The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *SOIL*, *6*(2), 565–578. https://doi.org/10.5194/soil-6-565-2020

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., et al. (2019). Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*, *352*, 251–267. https://doi.org/10.1016/j.geoderma.2019.06.016

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., et al. (2015). Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in Agronomy*, *132*, 139–159.

Oberholzer, S., Summerauer, L., Steffens, M., & Ifejika Speranza, C. (2024). Best performances of visible–near-infrared models in soils with little carbonate – a field study in Switzerland. *SOIL*, *10*(1), 231–249. https://doi.org/10.5194/soil-10-231-2024

Oldoni, H., Tavares, T. R., Brasco, T. L., Cherubin, M. R., de Carvalho, H. W. P., Magalhães, P. S. G., et al. (2024). Temporal evaluation of soil chemical quality using VNIR and XRF spectroscopies. *Soil and Tillage Research*, *240*, 106087. https://doi.org/10.1016/j.still.2024.106087

Pohlert, T. (2024). *PMCMRplus: Calculate pairwise multiple comparisons of mean rank sums extended*. https://CRAN.R-project.org/package=PMCMRplus

Quinlan, J. R. (1992). *Learning with continuous classes*. https://api.semanticscholar.org/CorpusID:1056674

Quinlan, J. R. (1993). Combining instance-based and model-based learning. *Proceedings of the Tenth International Conference on Machine Learning (ICML '93)*, 236–243. Morgan Kaufmann, University of Massachusetts, Amherst, MA. https://doi.org/10.1016/B978-1-55860-307-3.50037-X

R Core Team. (2024). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil properties. *Geoderma*, *195–196*, 268–279. https://doi.org/10.1016/j.geoderma.2012.12.014

Ramirez-Lopez, L., Wadoux, A., Franceschini, M., Terra, F., Marques, K. P. P., Sayão, V., et al. (2018). Robust soil mapping at the farm scale with vis–NIR spectroscopy. *European Journal of Soil Science*, *70*. https://doi.org/10.1111/ejss.12752

Renuka, K. P., Dharumarajan, S., Lalitha, M., Vasundhara, R., Kalaiselvi, B., & Sathish, A. (2026). Soil property prediction with vis-NIR spectroscopy in different depths: Testing the robustness of topsoil-derived models. *Journal of the Indian Society of Remote Sensing*, *54*(9), 3687–3699. https://doi.org/10.1007/s12524-026-02518-7

Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Sanderman, J., Gholizadeh, A., Pittaki-Chrysodonta, Z., Huang, J., Safanelli, J. L., & Ferguson, R. (2023). Transferability of a large mid-infrared soil spectral library between two Fourier-transform infrared spectrometers. *Soil Science Society of America Journal*, *87*, 586–599. https://doi.org/10.1002/saj2.20513

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Schmidinger, J., Barkov, V., Tavakoli, H., Correa, J., Ostermann, M., Atzmueller, M., et al. (2024). Which and how many soil sensors are ideal to predict key soil properties: A case study with seven sensors. *Geoderma*, *450*, 117017. https://doi.org/10.1016/j.geoderma.2024.117017

Schmidt, R. C. (1997). Managing Delphi Surveys Using Nonparametric Statistical Techniques. *Decision Sciences*, *28*(3), 763–774. https://doi.org/10.1111/j.1540-5915.1997.tb01330.x

Semella, S., Hutengs, C., Seidel, M., Ulrich, M., Schneider, B., Ortner, M., et al. (2022). Accuracy and reproducibility of laboratory diffuse reflectance measurements with portable VNIR and MIR spectrometers for predictive soil organic carbon modeling. *Sensors*, *22*(7), 2749. https://doi.org/10.3390/s22072749

Six, J., Elliott, E. T., & Paustian, K. (2000). Soil structure and soil organic matter II. A normalized stability index and the effect of mineralogy. *Soil Science Society of America Journal*, *64*(3), 1042–1049. https://doi.org/10.2136/sssaj2000.6431042x

Soil Science Division Staff. (2017). *Soil survey manual* (18th ed.). Government Printing Office, Washington, D.C., USA.

Song, J., Peng, C., Chen, Y., Zhao, S., Xu, X., & Xu, H. (2026). Province-scale mapping of cropland plough layer thickness using crop spectral response metrics and multi-source environmental covariates. *Agronomy*, *16*(16), 1520. https://doi.org/10.3390/agronomy16161520

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical,

chemical, and biological properties. *Applied Spectroscopy Reviews*, *49*(2), 139–186. https://doi.org/10.1080/05704928.2013.811081

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, *107*, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PLoS ONE*, *8*(6), e66409. https://doi.org/10.1371/journal.pone.0066409

Tilse, M., Bishop, T., Triantafilis, J., & Filippi, P. (2022). Mapping the impact of subsoil constraints on soil available water capacity and potential crop yield. *Crop & Pasture Science*, *73*, 636–651. https://doi.org/10.1071/CP21627

Tóth, G., Hermann, T., da Silva, M. R., & Montanarella, L. (2018). Monitoring soil for sustainable development and land degradation neutrality. *Environmental Monitoring and Assessment*, *57*. https://doi.org/10.1007/s10661-017-6415-3

Tümsavaş, Z., Tekin, Y., Ulusoy, Y., & Mouazen, A. (2019). Prediction and mapping of soil clay and sand contents using visible and near-infrared spectroscopy. *Biosystems Engineering*, *177*, 90–100. https://doi.org/10.1016/j.biosystemseng.2018.06.008

van Leeuwen, C. C. E., Mulder, V. L., & Heuvelink, G. B. M. (2026). The effect of modeller's choices and measurement errors on the performance of MIR soil spectroscopy models. *Geoderma*, *469*, 117804. https://doi.org/10.1016/j.geoderma.2026.117804

Verburg, P. H., Crossman, N., Ellis, E. C., Heinimann, A., Hostert, P., Mertz, O., et al. (2015). Land system science and sustainable development of the earth system: A global land project perspective. *Anthropocene*, *12*, 29–41. https://doi.org/10.1016/j.ancene.2015.09.004

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*(1–2), 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Ge, Y., et al. (2022). Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. *European Journal of Soil Science*, *73*(4), e13271. https://doi.org/10.1111/ejss.13271

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

Ward, K. J., Foerster, S., & Chabrillat, S. (2024). Estimating soil organic carbon using multitemporal PRISMA imaging spectroscopy data. *Geoderma*, *450*, 117025. https://doi.org/10.1016/j.geoderma.2024.117025

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, PLS Methods*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Yang, H., & Mouazen, A. M. (2012). Vis/Near- and Mid- Infrared spectroscopy for predicting soil N and C at a farm scale. In *Infrared Spectroscopy—Life and Biomedical Sciences*. IntechOpen. https://doi.org/10.5772/36393

Zayani, H., Fouad, Y., Michot, D., Kassouk, Z., Lili-Chabaane, Z., & Walter, C. (2023). Detecting the temporal trend of cultivated soil organic carbon content using visible near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *31*(5), 241–255. https://doi.org/10.1177/09670335231193113

Zhang, X., Xue, J., Xiao, Y., Shi, Z., & Chen, S. (2023). Towards optimal variable selection methods for soil property prediction using a regional soil vis-NIR spectral library. *Remote Sensing*, *15*(2). https://doi.org/10.3390/rs15020465

Zhou, Q., Ding, J., Ge, X., Li, K., Zhang, Z., & Gu, Y. (2023). Estimation of soil organic matter in the Ogan-Kuqa River Oasis, Northwest China, based on visible and near-infrared spectroscopy and machine learning. *Journal of Arid Land*, *15*(2), 191–204. https://doi.org/10.1007/s40333-023-0094-4

Zhou, Y., Chen, S., Hu, B., Ji, W., Li, S., Hong, Y., et al. (2022). Global soil salinity prediction by open soil vis-NIR spectral library. *Remote Sensing*, *14*(21). https://doi.org/10.3390/rs14215627

Zhou, Y., Jin, Y., Dai, L., Wang, Z., Wang, X., Zhao, J., et al. (2026). Optimal calibration sizes for predicting soil properties using vis-NIR spectra across scales. *European Journal of Soil Science*, *77*(2), e70316. https://doi.org/10.1111/ejss.70316

Table 1: Geostatistical indices and temporal variability of soil properties for topsoil (0–30 cm) and subsoil (30–60 cm) in 2011 and 2021. Shown are the Variance Ratio (VR), the Coefficient of Variation (CV in %), and the Nugget/Sill ratio as an indicator of spatial dependence. Mean, standard deviation (SD), minimum, maximum, skewness, and kurtosis are also reported for each property, depth, and year to characterize the underlying distribution. Significant differences in variance between 2011 and 2021 were evaluated using Levene's test ($p < 0.05$).

Depth	Variable	Year	Mean	SD	Min.	Max.	Skewn.	Kurt.	CV	Nugget/Sill	VR	Significance
Topsoil (0-30 cm)	pH_{H2O}	2011	7.2	0.4	6.2	7.9	-0.43	-0.6	5.55	0.05	2.7	***
		2021	7.4	0.65	6.0	8.6	0.01	-0.81	8.83	0.04		
	EC_{H2O}	2011	213.2	54.4	127	419	0.95	0.82	25.51	0.15	0.3	***
		2021	107.1	28.1	52	200	0.48	-0.07	26.27	0.31		
	$CaCO_3$	2011	0.44	0.29	0.14	1.88	2.32	6.28	66.93	0.42	1.1	n.s.
		2021	0.44	0.31	0.13	1.91	2.27	6.12	70.38	0.32		
	SOM	2011	4.22	0.32	2.68	4.98	-0.84	2.30	7.64	0.00	2.1	n.s.
		2021	4.03	0.47	2.57	7.5	2.48	18.1	11.61	0.62		
	clay	2011	7.71	0.93	5.52	10.3	0.18	0.00	12.02	0.44	0.9	n.s.
		2021	8.97	0.89	6.54	11.3	0.10	0.04	9.97	0.48		
	silt	2011	79.58	3.03	68.9	86.2	-0.64	1.11	3.81	0.39	0.8	n.s.
		2021	79.88	2.65	70.6	86.8	-0.38	1.04	3.31	0.41		
	sand	2011	12.71	3.16	5.99	24.1	0.76	0.89	24.85	0.47	0.8	n.s.
		2021	11.15	2.81	5	20.7	0.62	0.73	25.24	0.49		
Subsoil (30-60 cm)	pH_{H2O}	2011	7.7	0.38	6.4	8.9	-0.97	0.90	4.87	0.12	1.5	***
		2021	7.7	0.47	6.4	8.8	-0.81	0.03	6.06	0.12		
	EC_{H2O}	2011	216.4	76	89.3	502	1.05	0.90	35.11	0.11	0.2	***
		2021	130.4	32.1	65	284	1.04	2.64	24.60	0.40		
	$CaCO_3$	2011	1.16	1.64	0.15	8.63	2.73	7.16	141.60	0.30	0.4	*
		2021	0.85	1.02	0.14	7	3.21	11.8	121.00	0.43		
	SOM	2011	3.28	0.44	1.71	4.94	-0.11	1.79	13.46	0.40	1.0	n.s.
		2021	3.51	0.45	2.22	6.03	0.85	4.27	12.74	0.11		
	clay	2011	8.25	0.99	5.52	10.8	-0.48	-0.04	12.03	0.37	0.7	*
		2021	9.38	0.85	6.66	11.3	-0.33	0.22	9.11	0.56		
	silt	2011	79.78	6.21	51.3	88.3	-1.79	3.76	7.79	0.20	0.4	***
		2021	80.05	4.04	56.3	87.7	-1.90	6.45	5.04	0.00		
	sand	2011	11.97	6.62	3.73	42.8	1.87	3.93	55.28	0.16	0.5	***
		2021	10.57	4.49	4.41	37.1	1.99	6.41	42.50	0.00		

Table 2: Spatial autocorrelation analysis (Moran's I) of soil properties for topsoil (0–30 cm) and subsoil (30–60 cm) in 2011 and 2021. The results indicate the stability of spatial patterns and the resilience of the field's underlying soil architecture over a ten-year experimental period.

Layer	Parameter	2011		2021	
		Moran I	p-Value	Moran I	p-Value
Topsoil	pH_{H2O}	0.46	0.00	0.51	0.00
	EC_{H2O}	0.18	0.00	0.17	0.00
	$CaCO_3$	0.23	0.00	0.27	0.00
	SOM	0.09	0.00	0.13	0.00
	clay	0.03	2.26E-06	0.09	0.00
	silt	0.23	0.00	0.26	0.00
	sand	0.22	0.00	0.27	0.00
Subsoil	pH_{H2O}	0.46	0.00	0.47	0.00
	EC_{H2O}	0.15	0.00	0.14	0.00
	$CaCO_3$	0.29	0.00	0.19	0.00
	SOM	0.12	0.00	0.18	0.00
	clay	0.04	8.55E-09	0.09	0.00
	silt	0.17	0.00	0.16	0.00
	sand	0.17	0.00	0.16	0.00

Table 3: Spectral stability analysis based on PCA dispersion (Mean Distance to Centroid) for the years 2011 and 2021. The Stability Ratio (2021/2011) quantifies the resilience of the spectral point cloud size, demonstrating overall consistency in topsoil and natural stabilization (homogenization) in the subsoil. Per-sensor between-year similarity by the spectral angle mapper (SAM) and the mean absolute difference (MAD).

Layer	2011	2021	Stability Ratio	Data	SAM	MAD
Topsoil	38.75	39.16	1.01	Vis-NIR	2.24	0.027
				MIR	1.73	0.020
Subsoil	37.53	41.71	1.11	Vis-NIR	3.90	0.055
				MIR	1.99	0.027

Table 4: Friedman test **(χ^2)** results for the comparison of spectral datasets (Vis-NIR, MIR, synergistic dataset), stratified by depth.

Depth	χ^2	df	p	Kendall's W	Significant
Topsoil	10.57	2	0.005	0.76	Yes
Subsoil	7.14	2	0.028	0.51	Yes

Table 5: Nemenyi post-hoc pairwise comparisons between spectral datasets, stratified by depth.

Depth	Comparison	p	Significant
Topsoil	Synergistic vs MIR	0.009	Yes
Topsoil	Synergistic vs Vis-NIR	0.021	Yes
Topsoil	MIR vs Vis-NIR	0.961	No
Subsoil	Synergistic vs MIR	0.021	Yes
Subsoil	Synergistic vs Vis-NIR	0.375	No
Subsoil	MIR vs Vis-NIR	0.375	No

Table 6a: Friedman tests and Nemenyi post-hoc comparisons for algorithm selection on predictive accuracy (RPD by depth).

Depth	χ^2	df	p	Kendall's W	Significant	Post-hoc (Nemenyi)
Topsoil	3.71	2	0.156	0.27	No	ns (no significant pairwise differences)
Subsoil	11.14	2	0.004	0.80	Yes	Cubist > RF (p=0.004); Cubist > PLS (p=0.043); PLS–RF ns

Table 6b: Wilcoxon tests and Holm post-hoc comparisons for feature selection effect on predictive accuracy (RPD by depth).

Depth	Algorithm	Feature selection	p (Wilcoxon)	p (Holm-adjusted)	Significant
Topsoil	PLS	CARS	0.022	0.086	No
Topsoil	Cubist	RFE	0.452	0.961	No
Topsoil	RF	FRFS	0.374	0.961	No
Topsoil	RF	Boruta	0.320	0.961	No
Subsoil	PLS	CARS	0.708	1.000	No
Subsoil	Cubist	RFE	0.137	0.412	No
Subsoil	RF	FRFS	0.973	1.000	No
Subsoil	RF	Boruta	0.002	0.009	Yes

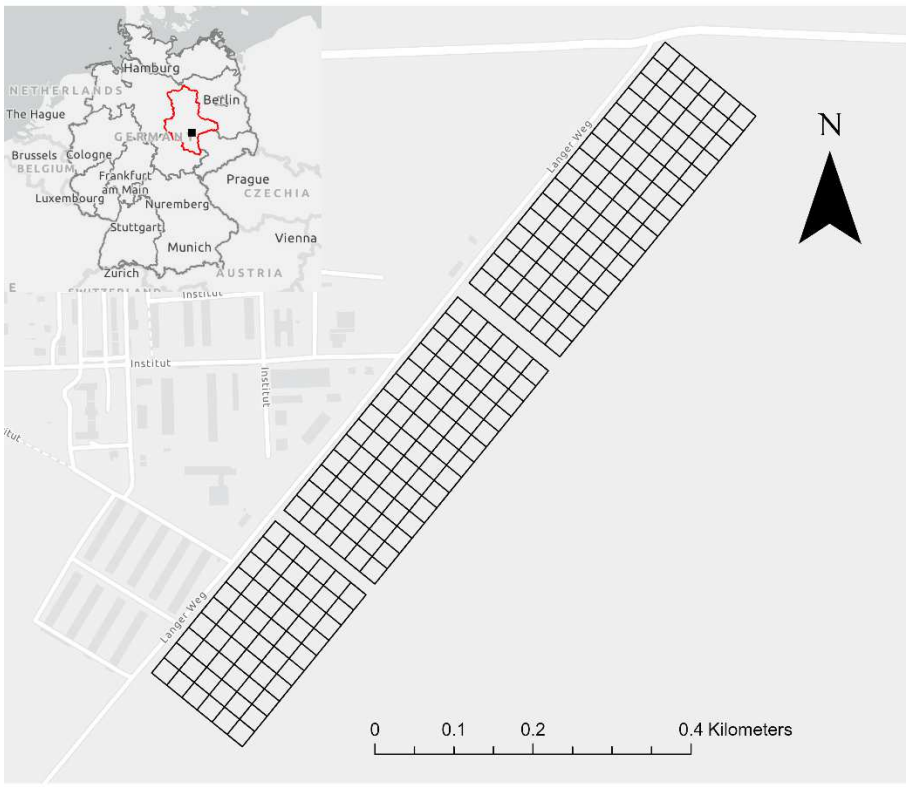


Fig. 1: Geographical location and layout of the study area. The inset shows the location of Saxony-Anhalt (red outline) and Crop Research Unit Merbitz (black dot). The main figure shows long-term experimental field, subdivided into the sampling grid used for soil collection.

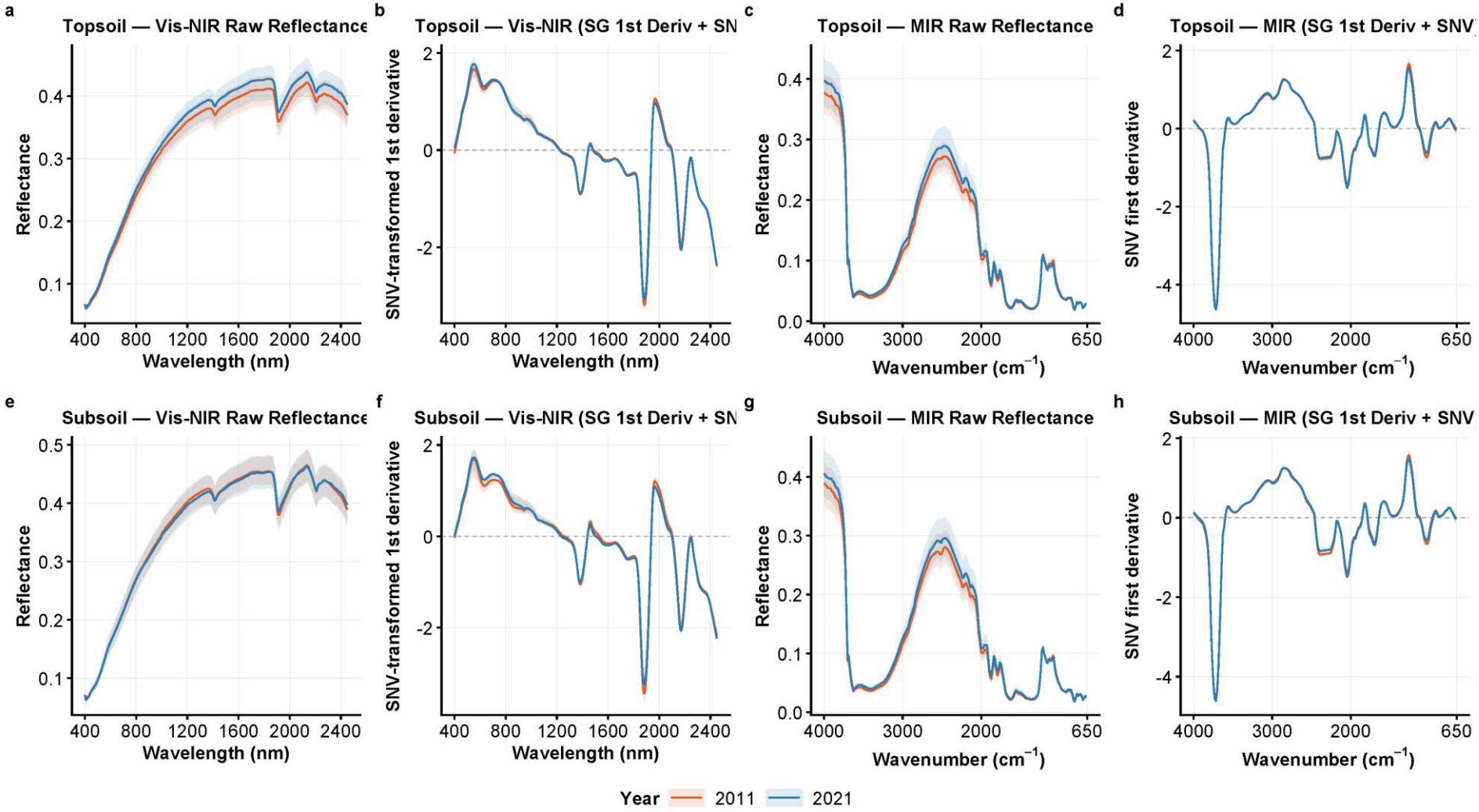


Fig. 2: Mean reflectance spectra of Vis-NIR and MIR for topsoil (a–d) and subsoil (e–h), scanned with RS-3500 (Vis-NIR) and Agilent 4300 handheld FTIR device (MIR). Within each depth, panels show Vis-NIR raw (a, e) and after Savitzky–Golay first-derivative + SNV transformation (b, f), and MIR raw (c, g) and after the same transformation (d, h). Solid lines denote the mean spectrum for each sampling year (2011, orange; 2021, blue); shaded bands represent ±1 standard deviation.

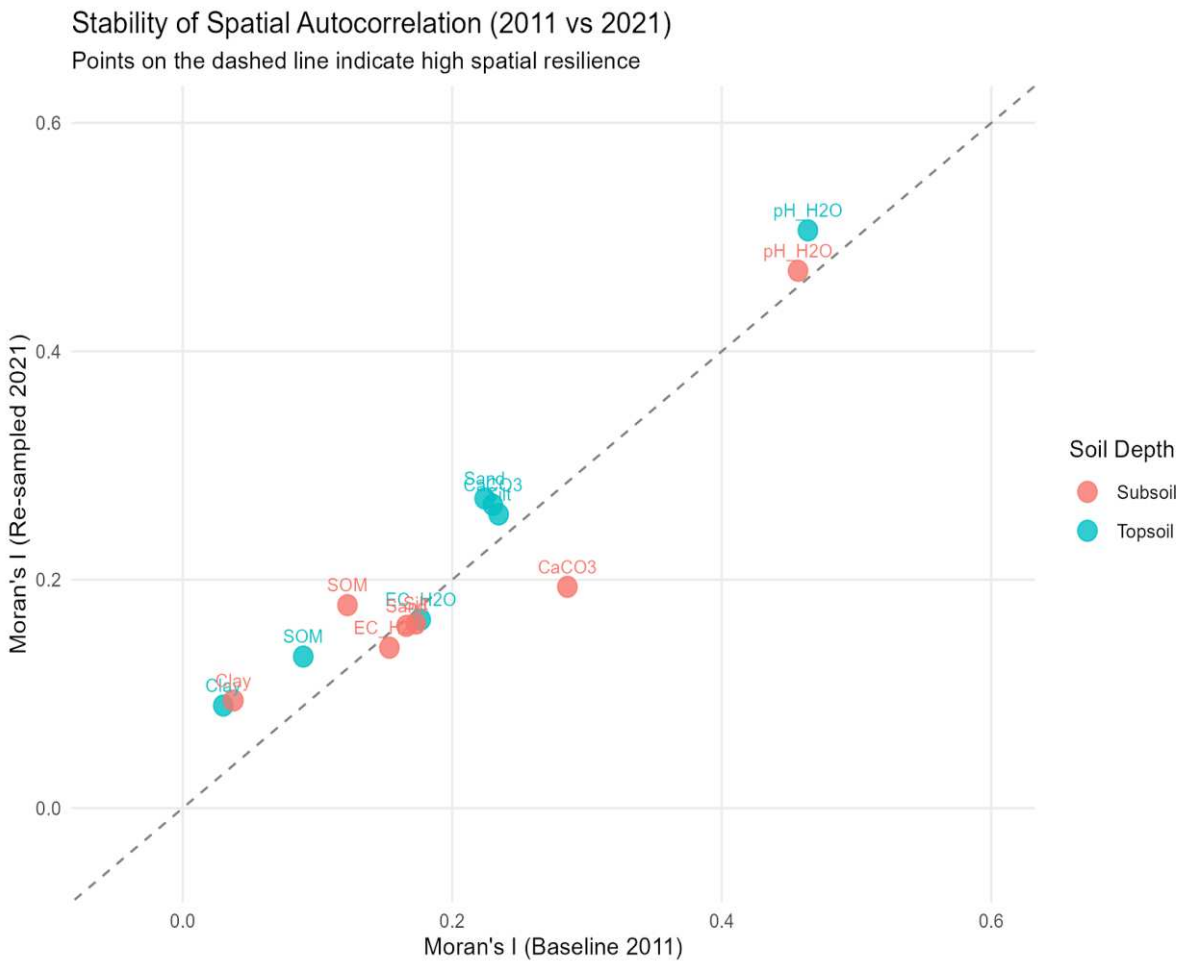
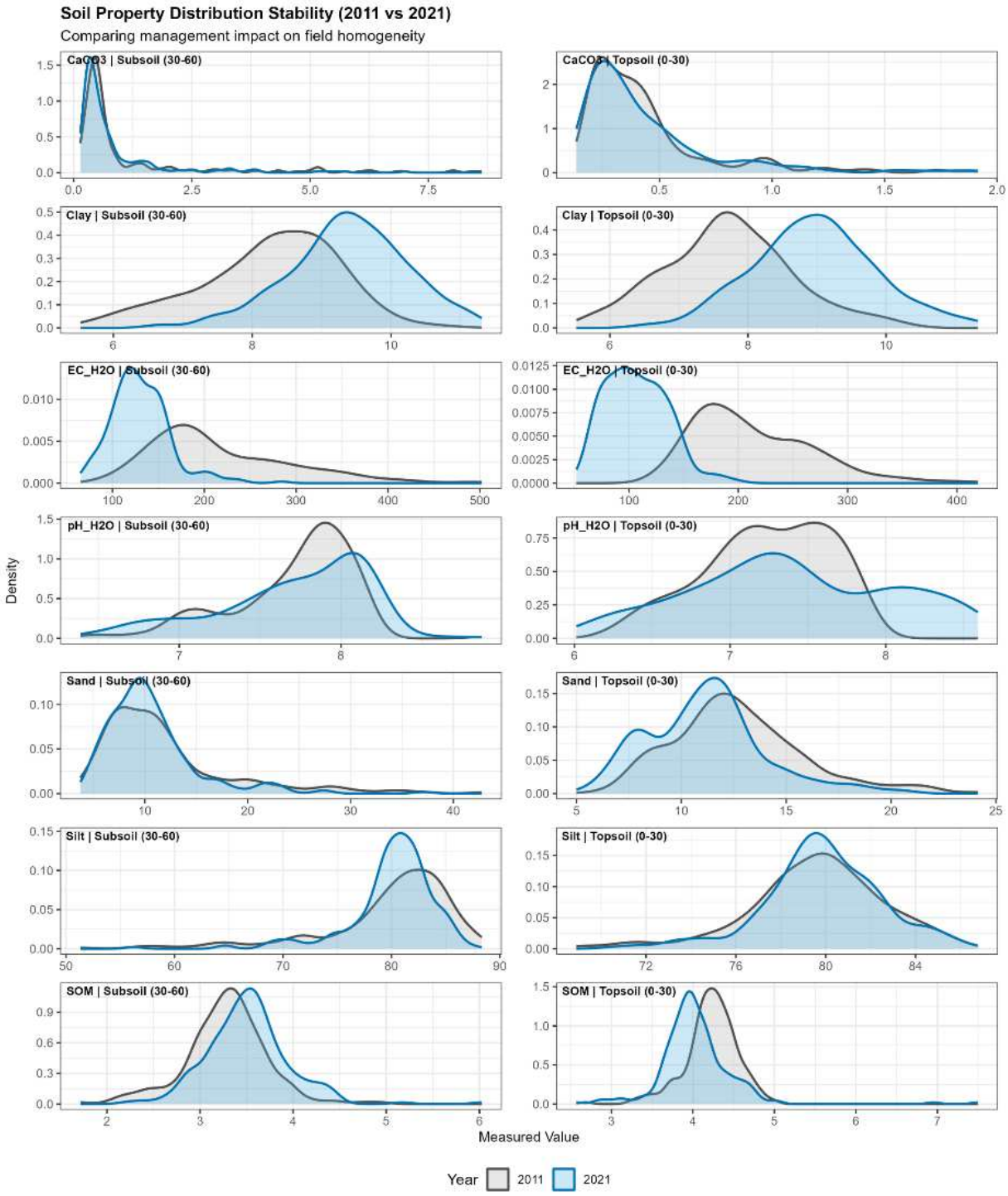


Fig. 3: Stability of spatial autocorrelation (Moran's I) for soil properties over a decade (2011 vs. 2021). Comparison of spatial structures in topsoil (0–30 cm) and subsoil (30–60 cm) at the baseline (2011) and after ten years of experimental field management (2021). Points located on or near the dashed 1:1 line indicate high spatial resilience of the soil architecture. **1**While physical properties (texture, SOM) maintain stable, low levels of autocorrelation, chemical parameters (pH) exhibit stronger but highly consistent spatial patterns over time.



1267

1268 **Fig. 4:** Comparative density distribution of soil properties (2011 vs. 2021) for topsoil (0–30 cm) and subsoil (30–
1269 60 cm). The plots illustrate the stability of field homogeneity over ten years of experimental management. While
1270 physical properties (texture: clay, silt, sand; organic matter: SOM) show highly consistent distribution patterns,
1271 chemical parameters (pH, EC) exhibit noticeable shifts (temporal drift) and increased variance in the topsoil. This
1272 reflects the impact of targeted experimental treatments (e.g., fertilization, liming), whereas the fundamental
1273 physical site characteristics remain resilient.

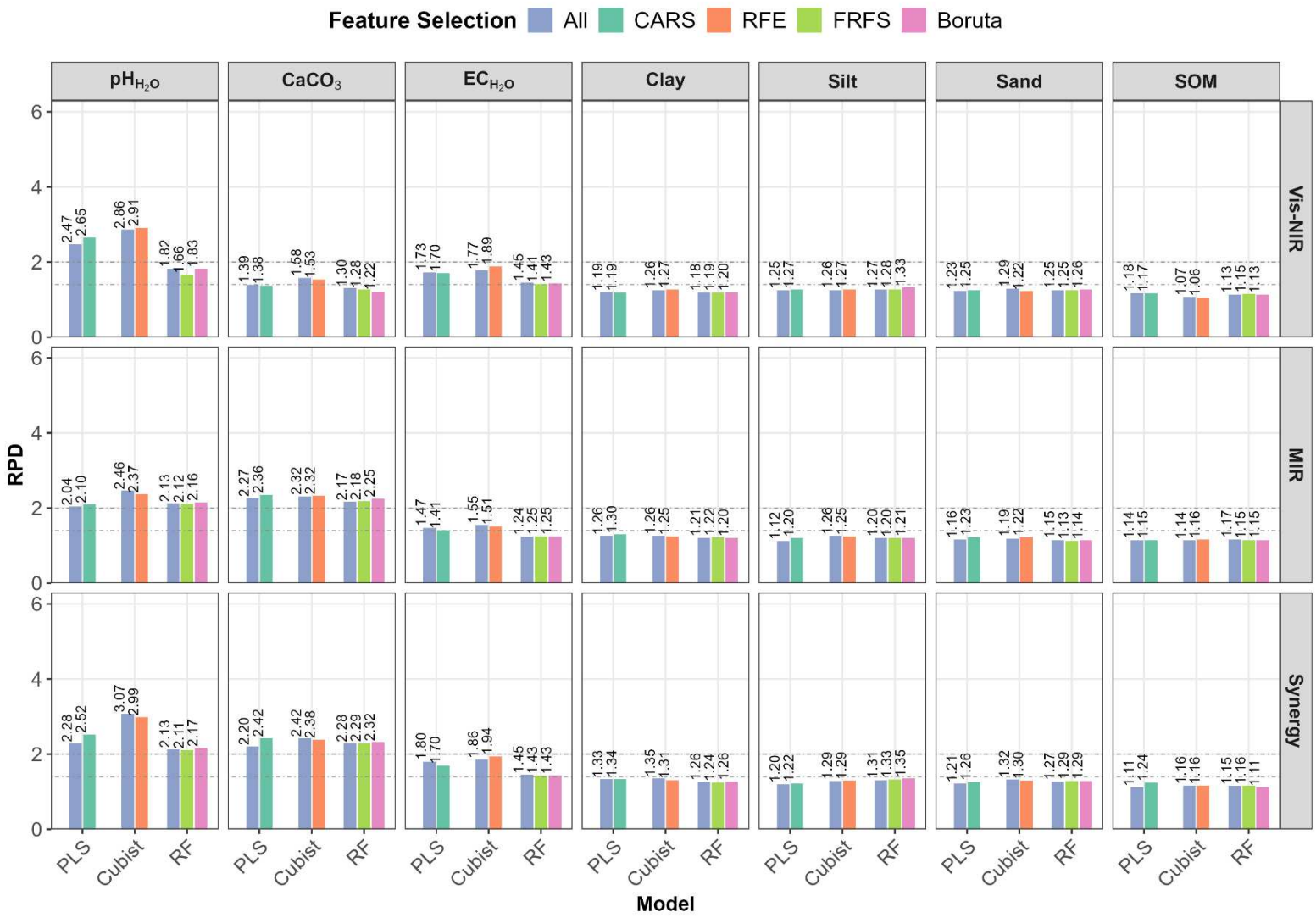


Fig. 5: Ratio of performance to deviation (RPD) for the prediction of seven soil properties in the topsoil dataset under the 70:30 split. Each machine learning algorithm was paired only with its applicable feature selection methods. Dash lines mark the RPD thresholds for fair (1.4) and good (2.0) prediction

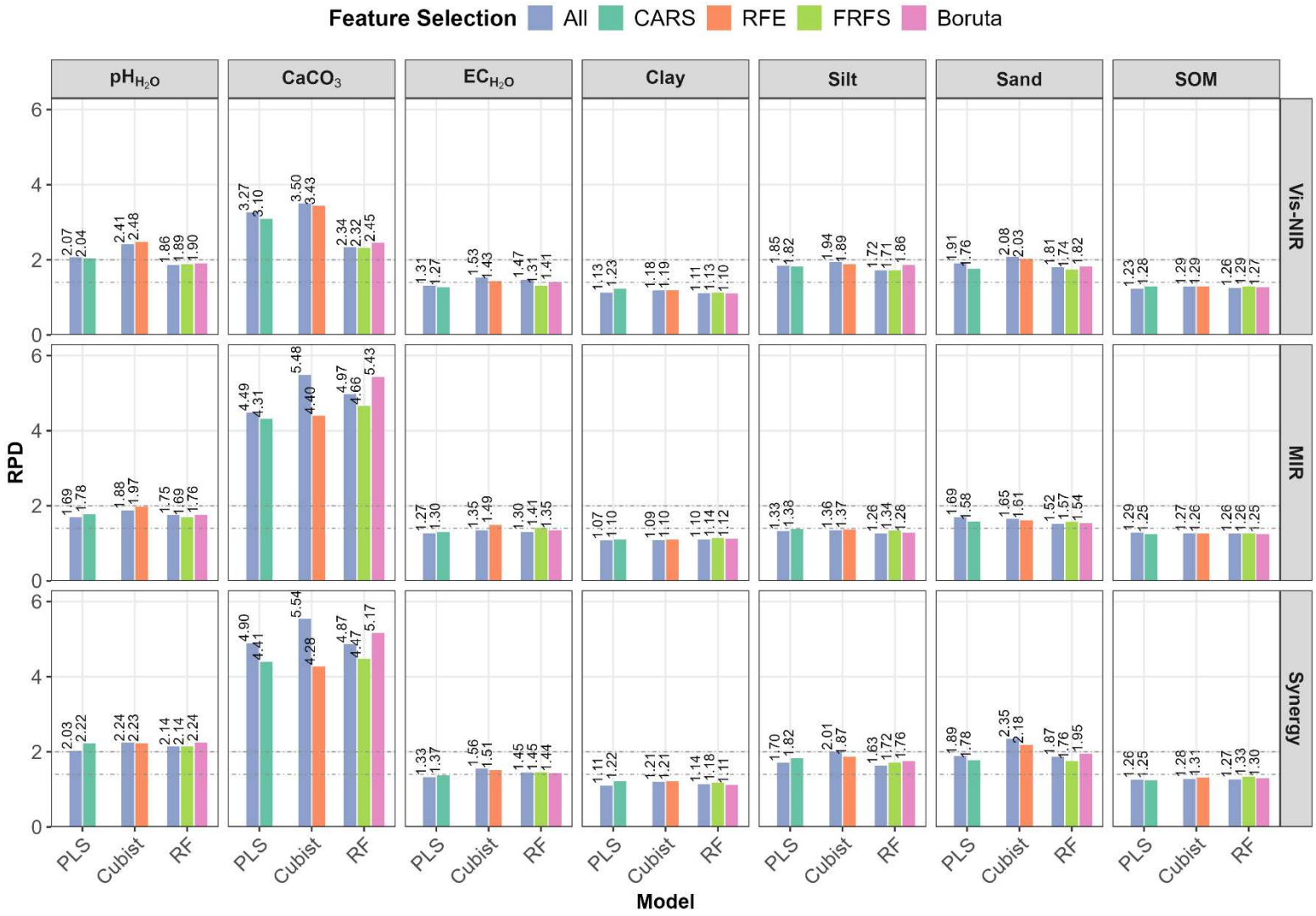


Fig. 6: Ratio of performance to deviation (RPD) for the prediction of seven soil properties in the subsoil dataset under the 70:30 split. Each machine learning algorithm was paired only with its applicable feature selection methods. Dash lines mark the RPD thresholds for fair (1.4) and good (2.0) prediction

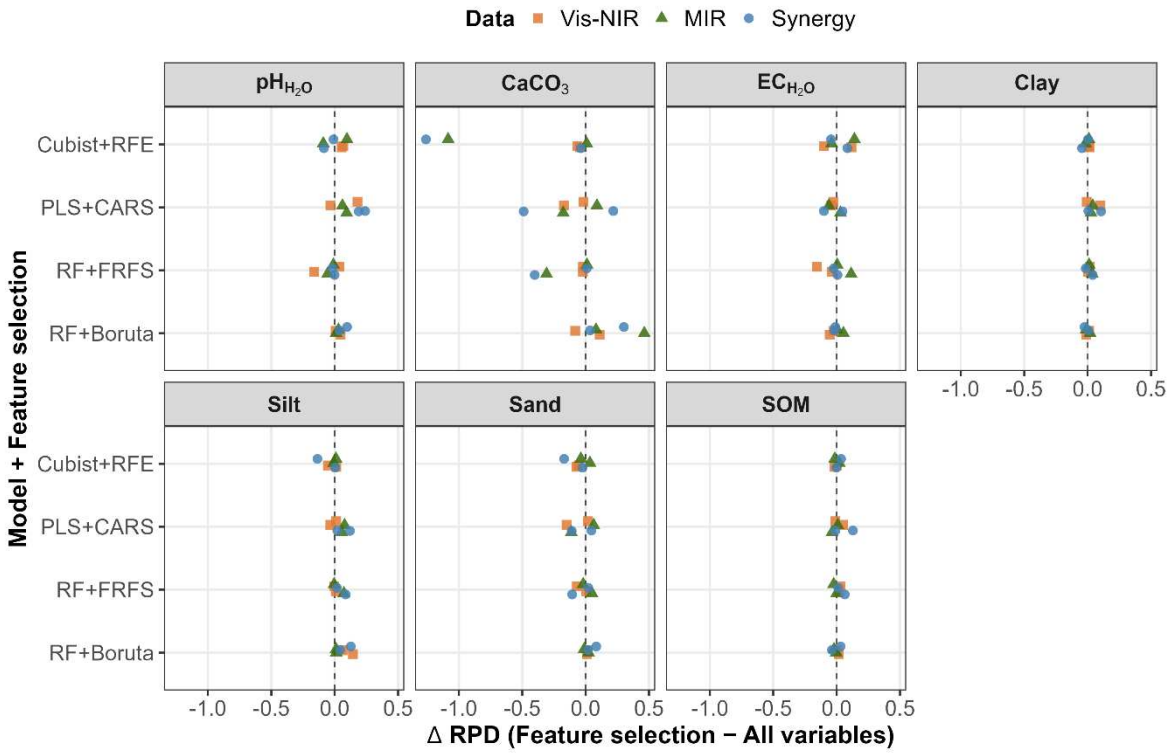


Fig. 7: Property-wise feature-selection advantage over the full variable set, expressed as ΔRPD (feature selection − all variables), for each algorithm–feature-selection pairing (Cubist–RFE, PLS–CARS, RF–FRFS, RF–Boruta) across the seven soil properties. The dashed vertical line marks zero.

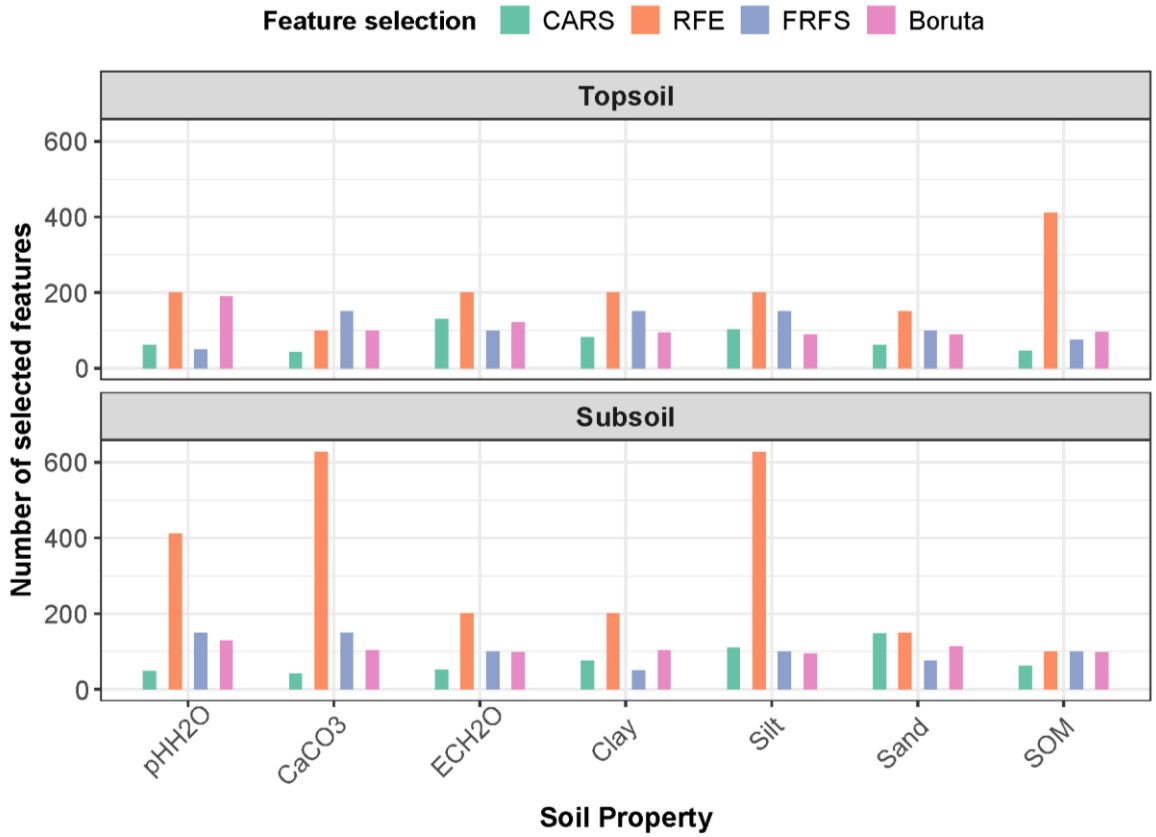


Fig. 8: Number of variables retained by each feature-selection method (CARS, RFE, FRFS, and Boruta) for each soil property and depth.

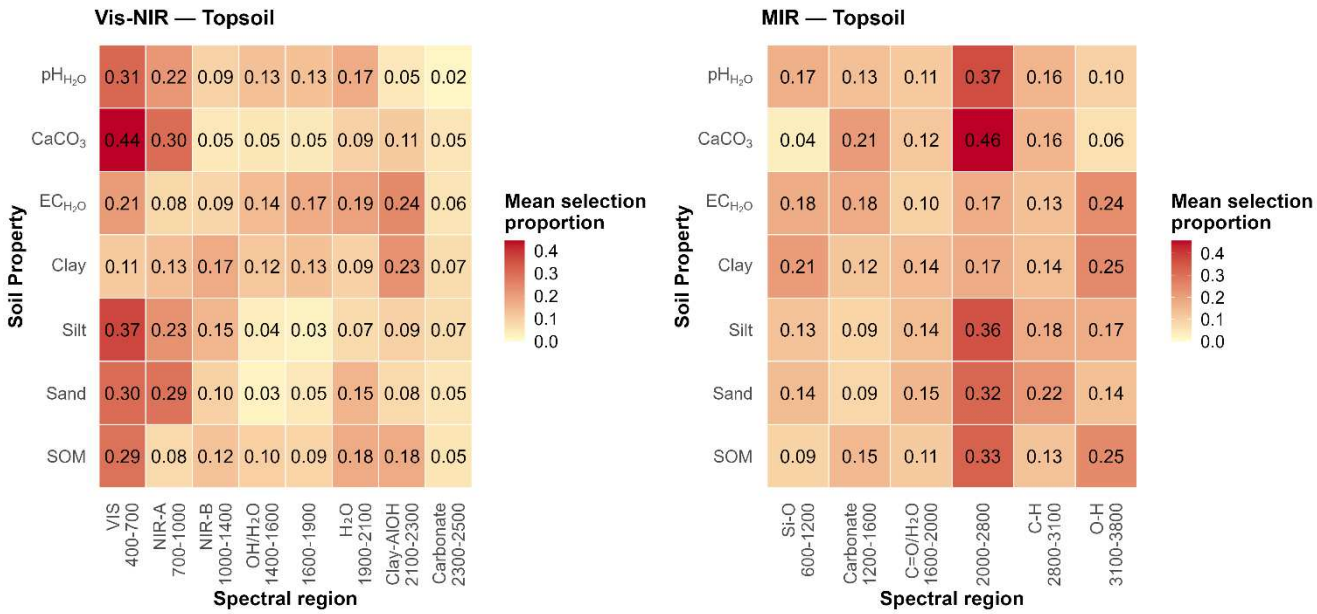


Fig. 9: Spectral region selection proportions per soil property for the topsoil dataset, for the Vis-NIR and MIR ranges. Cell values give the proportion of each method's selected bands falling within a spectral region, averaged across the four feature-selection methods (CARS, RFE, FRFS, Boruta); darker shading indicates a region more consistently selected. Regions are labelled by their associated absorption features defined according to standard soil spectroscopy literature (covering electronic transitions in the Vis-NIR and fundamental molecular vibrations in the MIR) (Comstock et al., 2019; Soriano-Disla et al., 2014).

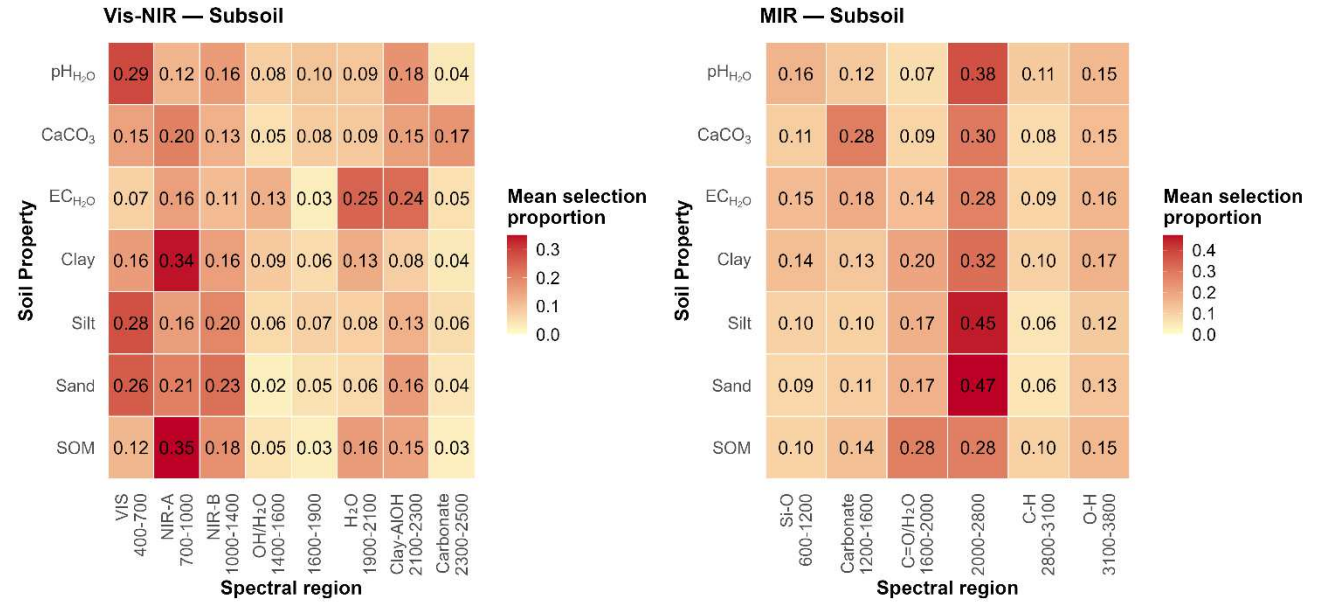


Fig. 10: Spectral region selection proportions per soil property for the subsoil dataset, for the Vis-NIR and MIR ranges. Cell values give the proportion of each method's selected bands falling within a spectral region, averaged across the four feature-selection methods (CARS, RFE, FRFS, Boruta); darker shading indicates a region more consistently selected. Regions are labelled by their associated absorption features defined according to standard soil spectroscopy literature (covering electronic transitions in the Vis-NIR and fundamental molecular vibrations in the MIR) (Comstock et al., 2019; Soriano-Disla et al., 2014).

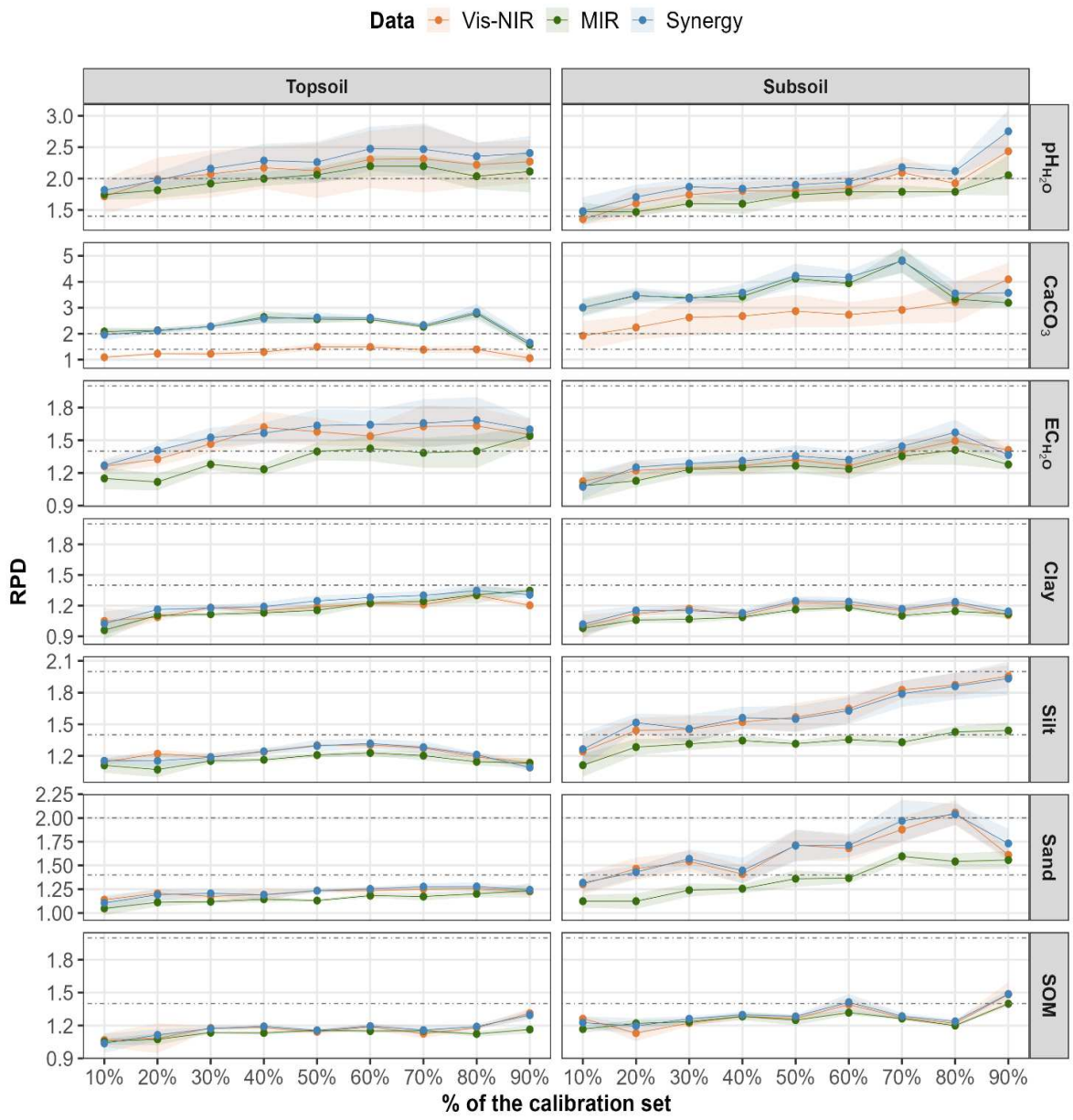


Fig. 11: Learning curves showing RPD as a function of calibration set size (10–90% of samples used for training) for each soil property and across top and subsoil depths. Coloured lines indicate the spectral dataset (Vis-NIR, MIR, and Synergy); shaded regions show ±1 standard deviation across all seven model–feature-selection combinations. Horizontal dashed lines indicate RPD = 1.4 (fair) and 2.0 (good).

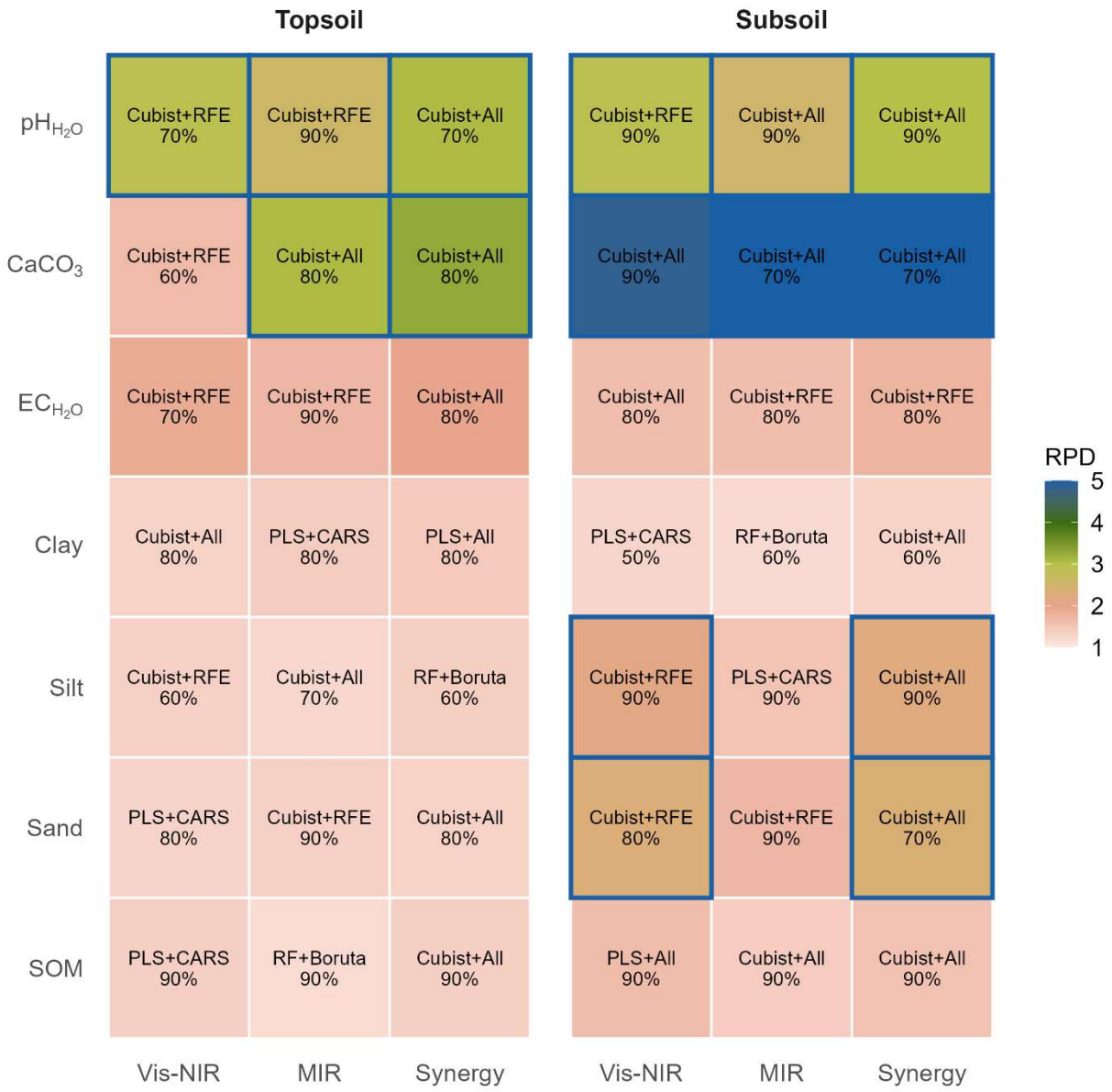


Fig. 12: Best achievable RPD for each soil property, spectral dataset, and depth. Each cell shows the best model and feature-selection combination and the calibration proportion at which it was obtained. Blue borders indicate predictions at RPD ≥ 2.0.

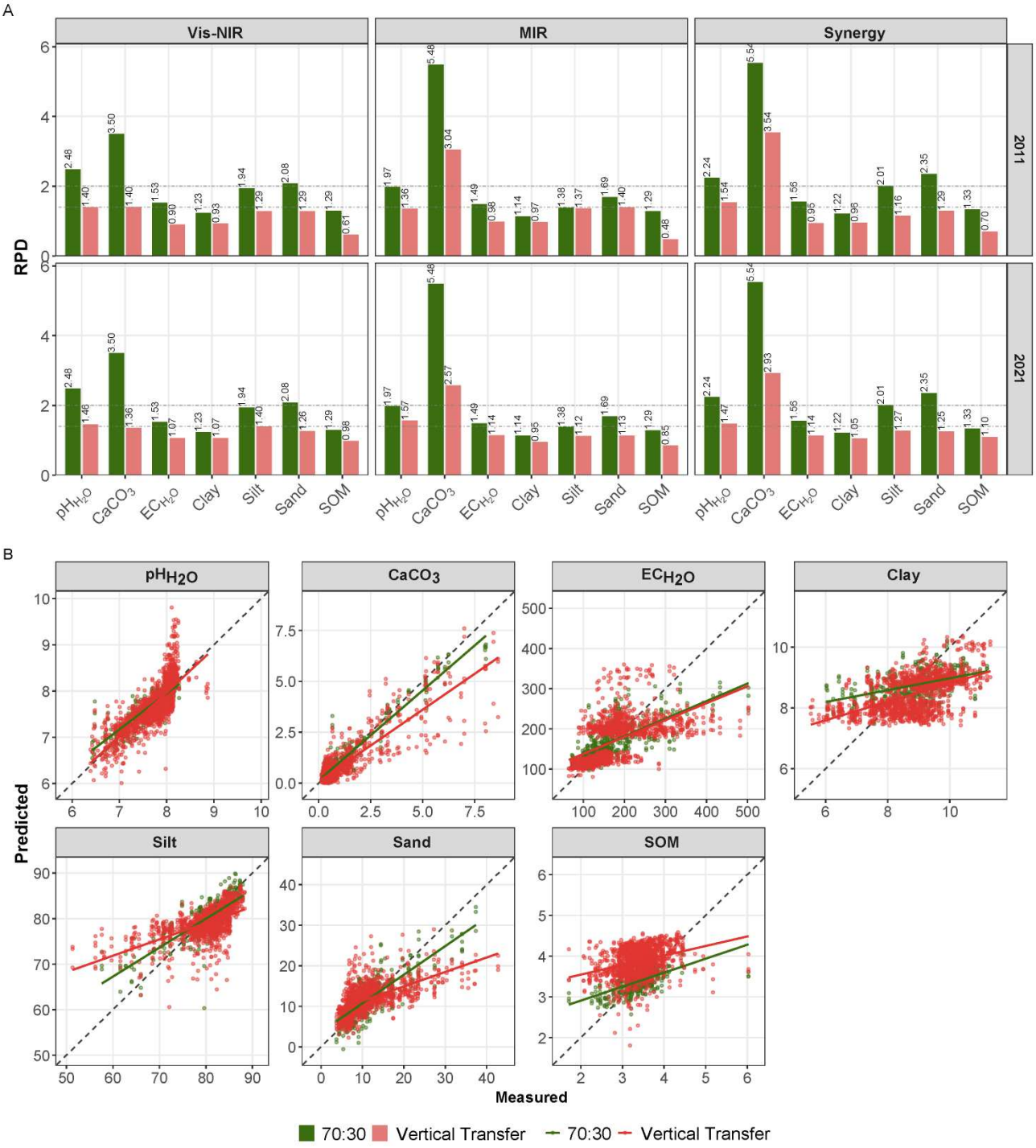


Fig. 13: Vertical transfer (topsoil → subsoil) of prediction models. **(A)** RPD of the best-performing model for each soil property and sensor, comparing the baseline (70:30) with the transfer performance, shown separately for 2011 and 2021; dash lines mark the RPD thresholds for fair (1.4) and good (2.0) prediction. **(B)** Measured versus predicted values of the best model from 70:30 and vertical transfer models for each property; the dashed and soil lines represent the 1:1 and regression lines, respectively.

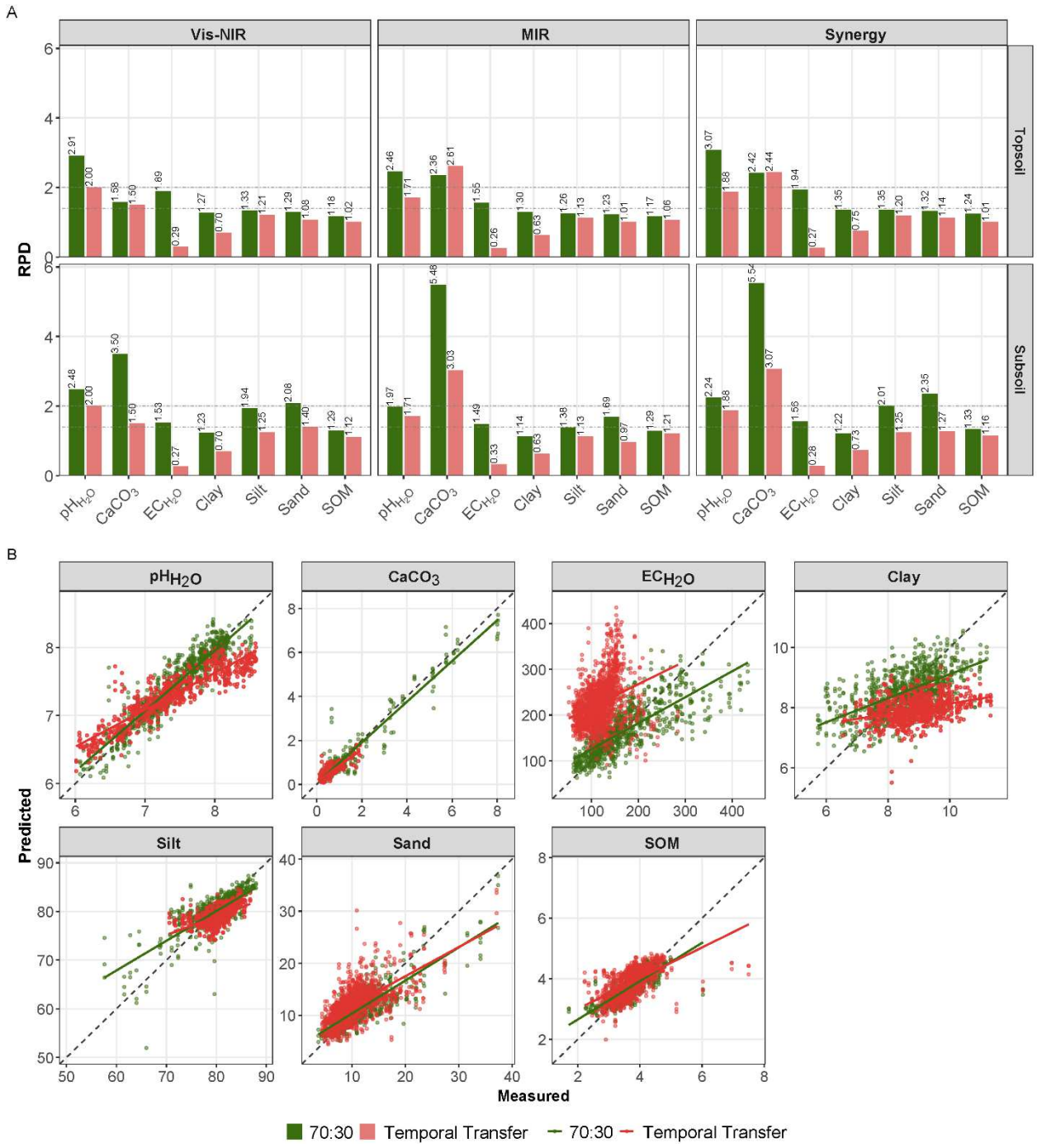


Fig. 14: Temporal transfer (2011 → 2021) of prediction models. **(A)** RPD of the best-performing model for each soil property and sensor, comparing the baseline (70:30) with the transfer performance, shown separately for topsoil and subsoil; dash lines mark the RPD thresholds for fair (1.4) and good (2.0) prediction. **(B)** Measured versus predicted values of the best model from 70:30 and vertical transfer models for each property; the dashed and soil lines represent the 1:1 and regression lines, respectively.

Appendix

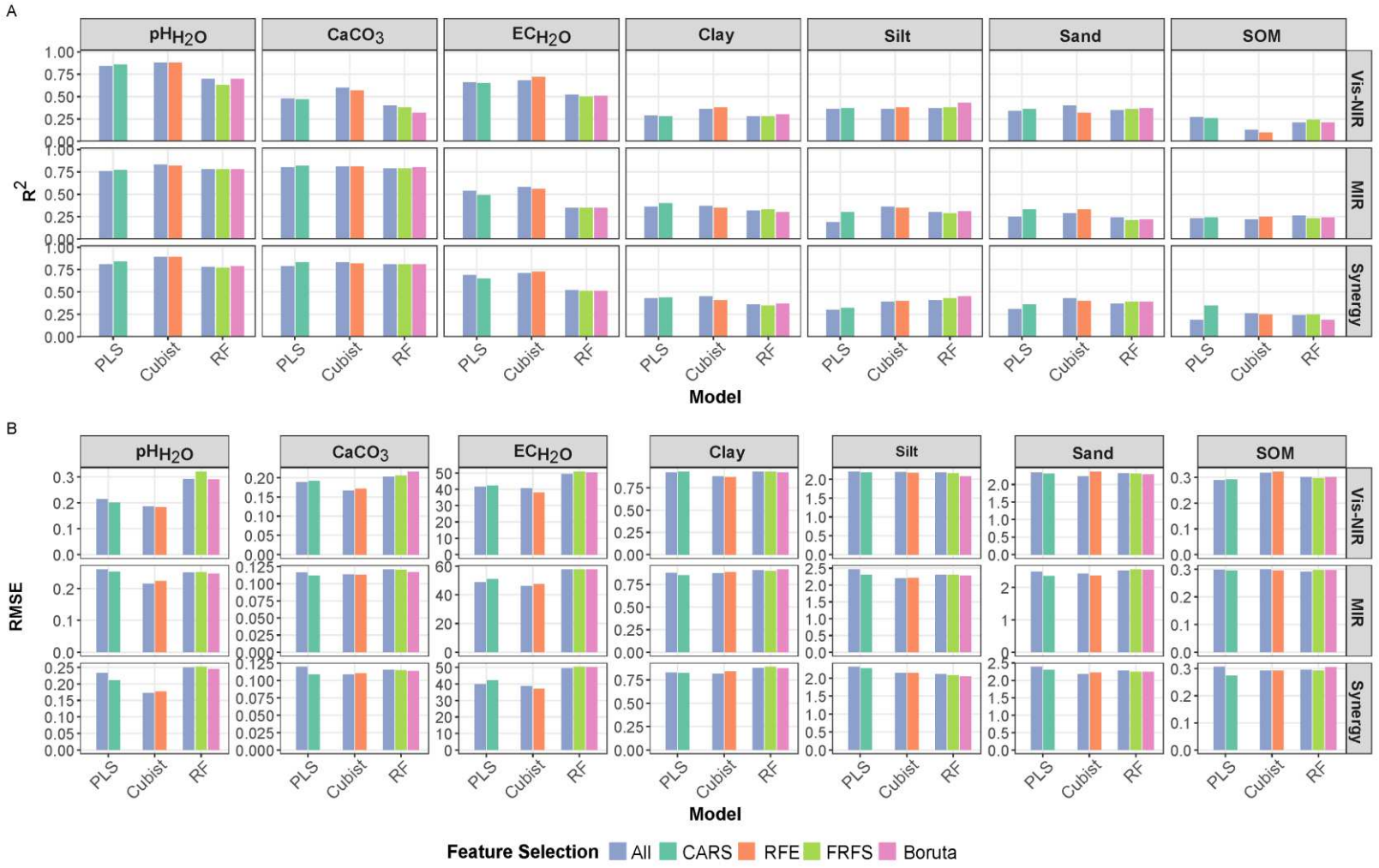


Fig. A1: (A) Coefficient of determination (R^2) for the prediction and **(B)** root mean square error (RMSE) of seven soil properties in the topsoil dataset under the 70:30 split. Each machine learning algorithm was paired only with its applicable feature selection methods.

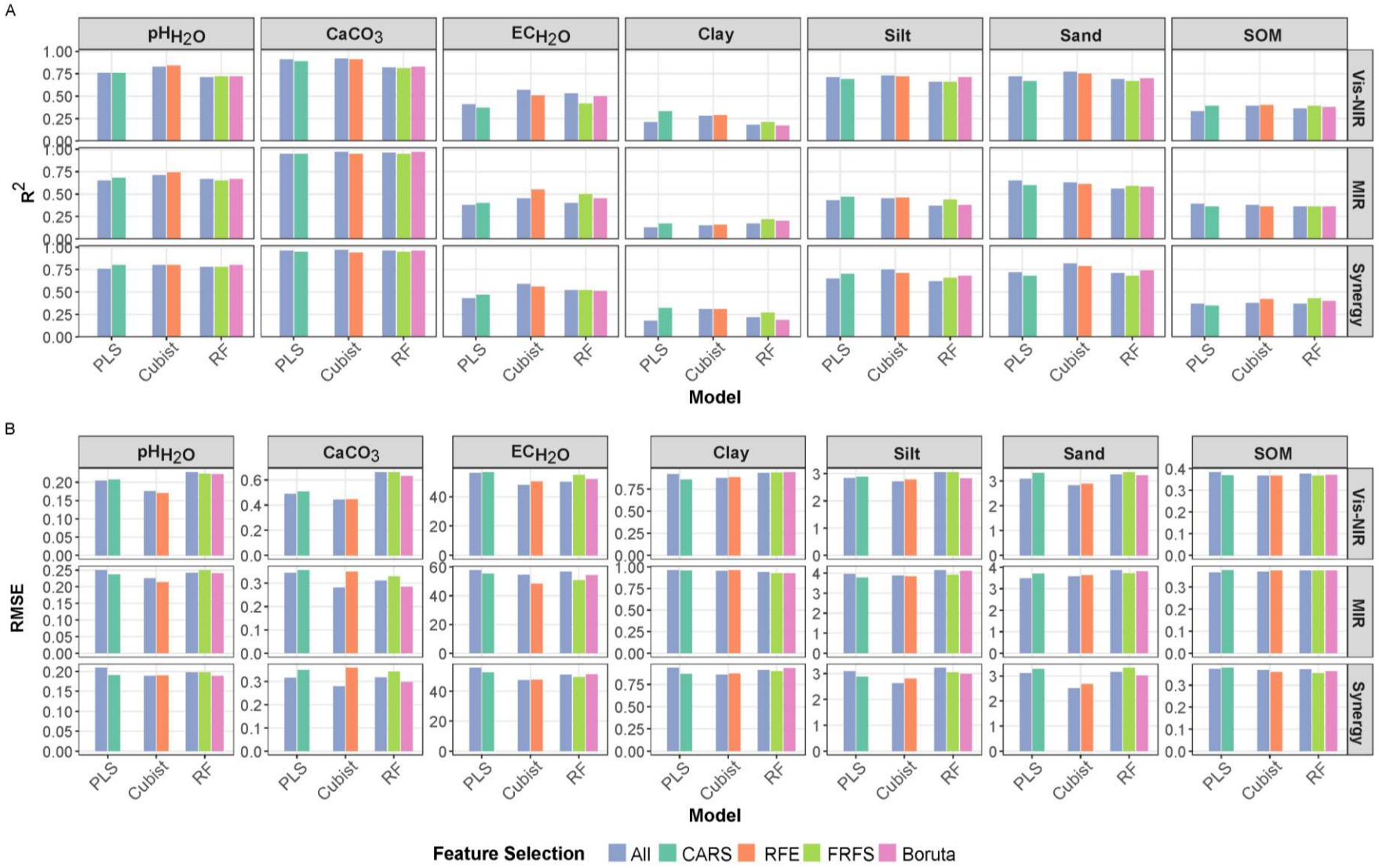


Fig. A2: (A) Coefficient of determination (R^2) for the prediction and **(B)** root mean square error (RMSE) of seven soil properties in the subsoil dataset under the 70:30 split. Each machine learning algorithm was paired only with its applicable feature selection methods.

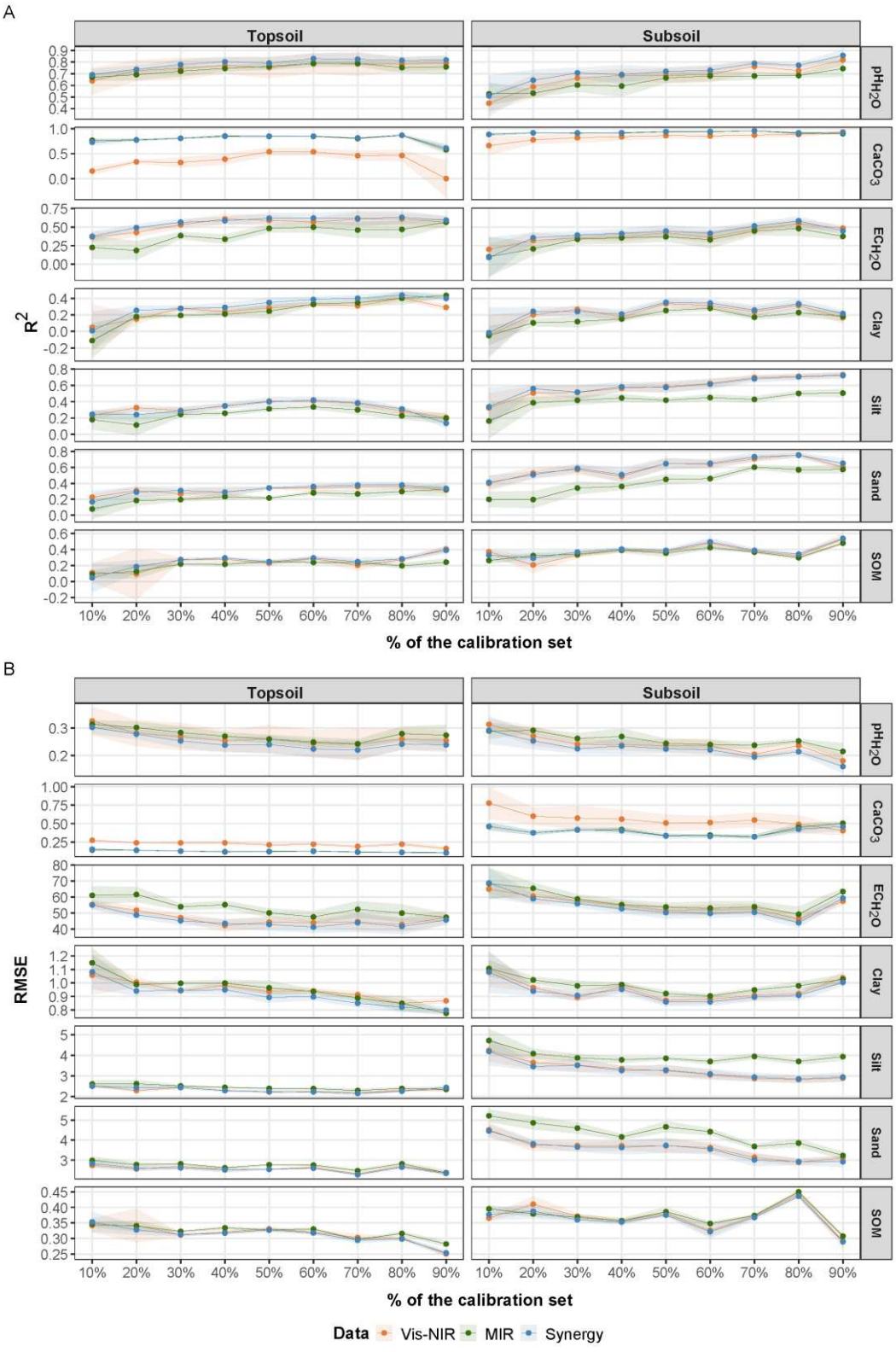


Fig. A3: Learning curves showing R^2 **(A)** and RMSE **(B)** as a function of calibration set size (10–90% of samples used for training) for each soil property and across topsoil and subsoil depths. Coloured lines indicate the spectral dataset (Vis-NIR, MIR, and Synergy); shaded regions show ±1 standard deviation across all seven model–feature-selection combinations

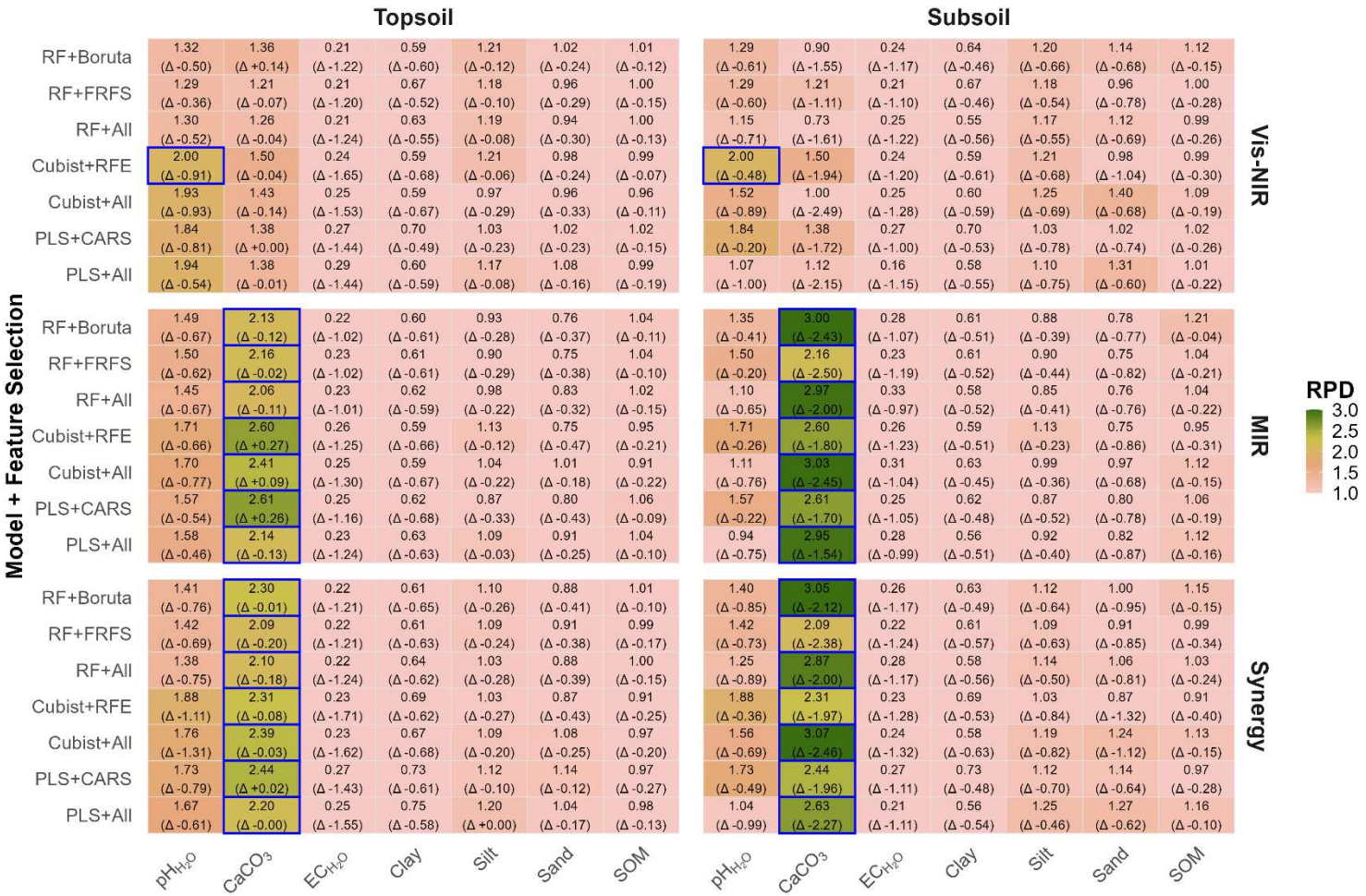


Fig. A4: Temporal transfer (2011 to 2021) RPD for all model–feature-selection combinations by soil property, spectral dataset (Vis-NIR, MIR, Synergy) and depth. Δ represents the change from the 70:30 baseline data; negative values indicate reduced accuracy under transfer. blue outlines mark RPD ≥ 2.



Fig. A5: Vertical transfer (topsoil to subsoil) RPD for all model–feature-selection combinations by soil property, spectral dataset (Vis-NIR, MIR, Synergy) and depth. Δ represents the change from the 70:30 baseline data; negative values indicate reduced accuracy under transfer. blue outlines mark RPD ≥ 2

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PoehlitzSpectrometrySupplementaryMaterialPAG.docx