Best performances of visible-near infrared models in soils with little carbonate - a field study in Switzerland

Simon Oberholzer[1], Laura Summerauer[2], Markus Steffens[1,3], Chinwe Ifejika Speranza[1]

[1]Institute of Geography, University Bern, 3012 Bern, Switzerland
[2]Department of Environmental System Science, ETH Zürich, 8092 Zürich, Switzerland
[3]Department of Soil Science, Research Institute of Organic Agriculture, 5070 Frick, Switzerland

Correspondence to: Laura Summerauer (laura.summerauer@usys.ethz.ch)

Abstract. Conventional laboratory analysis of soil properties is often expensive and requires much time if various soil properties are to be measured. Visual and near infrared (vis–NIR) spectroscopy offers a complementary and cost-efficient way to gain a wide variety of soil information in high spatial and temporal resolution. Yet, applying vis-NIR spectroscopy requires confidence in the prediction accuracy of the infrared models. In this study we used soil data from six agricultural fields in Eastern Switzerland and calibrated i) field-specific (local) models and ii) general models (combining all fields) for soil organic carbon (SOC), permanganate oxidizable carbon (POXC), total nitrogen (N), total carbon (C) and pH using partial least squares regression. The 30 local models showed a ratio of performance to deviation (RPD) between 1.14 and 5.27 and the root mean square errors (RMSE) were between 1.07 and 2.43 g kg^{-1} for SOC, between 0.03 and 0.07 g kg^{-1} for POXC, between 0.09 and 0.14 g kg^{-1} for total N, between 1.29 and 2.63 g kg^{-1} for total C and between 0.04 and 0.19 for pH. Two fields with high carbonate content and poor correlation between the target properties were responsible for six local models with a low performance (RPD < 2). Analysis of variable importance in projection as well as correlations between spectral variables and target soil properties confirmed that high carbonate content masked absorption features for SOC. Field sites with low carbonate content can be combined to general models with only limited loss in prediction accuracy compared to the field-specific models. On the other hand, for fields with high carbonate contents the prediction accuracy substantially decreased in general models. Whether the combination of soils with high carbonate contents in one prediction model leads to satisfying prediction accuracies needs further investigation.

1 Introduction

The application of spectroscopy in the visible and near infrared (vis–NIR) range is increasing in soil science and related disciplines with the main objective to gain information on soil properties of more samples at lower costs than with conventional laboratory methods. With a larger sample size, the spatial or temporal resolution can be increased which allows conclusions about the within-field or within-farm variability but might potentially also increase the statistical power in agricultural experiments (Greenberg et al., 2022). Despite its tendency to be less accurate compared to mid infrared (MIR) spectroscopy,

30   vis–NIR spectroscopy is widely applied because of less sample preparation, lower costs and generally easier portability (Soriano-Disla et al., 2014).

On-site vis–NIR measurements are therefore feasible but laboratory measurements with dried and sieved soil samples have so far shown higher accuracy (Allory et al., 2019; Hutengs et al., 2019). In particular, soil properties related to soil organic matter can be estimated appropriately by laboratory vis–NIR spectroscopy (Angelopoulou et al., 2020). In most cases, the focus is to

35   provide soil information over large areas (e. g. soil maps) where a high sample number is present and only a moderate prediction accuracy is needed. Hence, large-scale spectral libraries have been developed, to further reduce the need for wet chemistry data. Due to the high complexity within spectral libraries, the application of a general model to a local context leads to high prediction errors. Recent research shows that the localization of these infrared models substantially improves the predictive performance in a local context, for example by spiking (Brown, 2007; Li et al., 2020; Ng et al., 2022; Seidel et al.,

40   2019; Wetterlind and Stenberg, 2010; Zhao et al., 2021), memory-based learning (Ramirez-Lopez et al., 2013), resampling algorithms (Lobsey et al., 2017) or deep learning (Shen et al., 2022). However, for analyzing small-scale variability (field or farm level), a local model is often still the best choice because of its low prediction errors. Theoretically, developing local models is supported by the finding that in the vis–NIR range, spectral features that influence specific soil properties vary strongly between different datasets, which makes highly heterogenous large datasets prone to insufficient model performance

45   (Angelopoulou et al., 2020; Grunwald et al., 2018). The development of local spectral models has the main purpose to cope with a large sample size at the local scale, but such local models have no utility beyond the analysis of the specific local dataset. Spectral vis–NIR models developed from local datasets showed a very high variability in model performance ranging from excellent models (Breure et al., 2022; Seidel et al., 2019) to those with relatively poor model performance (Camargo et al., 2022; Kuang and Mouazen, 2011). The reasons for these different performances of local models are understudied and remain

50   unclear. Among many different possible modelling approaches including support vector machine regression, artificial neural networks, cubist and random forest, partial least square regression (PLSR) is the most frequently used model type to build spectral models with small datasets (Alomar et al., 2021; Zhao et al., 2021).

The number of samples is crucial for local models because often, only a limited number of samples with reference laboratory data are available. Kuang and Mouazen (2012) showed that local models improve with increasing numbers of calibration

55   samples and that a sample size of at least 50 provides accurate prediction models. Some studies thus combined multiple target sites and develop a general model by combining all the local datasets to reach a larger sample size and potentially better model performance (Kuang and Mouazen, 2011; Singh et al., 2022). In these studies, the general model showed an intermediate performance, and the general prediction error was between the best and the poorest performing local model. However, these studies only calculated the overall prediction error of the general model and therefore it is not clear if the prediction on target

60   sites with poorly performing local models could be improved by applying a general model.

For vis–NIR spectroscopy application at local scales, it is therefore very difficult to estimate the measurement accuracy for the predicted samples beforehand. This uncertainty is probably the main reason that hampers the application of vis–NIR spectroscopy because researchers prefer to rely on conventional lab measurements with a smaller sample size (and smaller

spatial resolution) where the measurement accuracy is known before sampling and measurements are conducted. Applying
65   spectroscopy at field or farm scale thus bears the risk that the measurement accuracy (RMSE) may be beyond the tolerable
threshold, which might then question a whole project. Thus, in this paper, we analyze the performance of field-specific (local)
spectral models of a field experiment conducted in six fields in Eastern Switzerland and that of a general model combining the
data from all six fields to ascertain their influencing factors. We ask:

1.   To what extent do the prediction errors of local spectral models differ from the lab measurement error?
70   2.   Does a general model that includes several target sites improve the prediction on a target site with a poor local model
performance?
3.   How do field and soil characteristics (e. g. field size, soil texture, carbonate content, correlations of soil properties)
of the target site relate to the performance of spectral models?

By answering these questions, we want to provide insights for estimations of prediction accuracies for vis–NIR studies at the
75   local scale with the objective to support decision making during the development of a sampling design and planning of
laboratory reference measurements for subsequent calibration modeling.

2 Methods

2.1 Datasets from a cover cropping experiment on six field sites

We used datasets from six fields (A, B, C, D, E, F) of a cover cropping experiment in the Canton of Thurgau, Eastern
80   Switzerland (paper in preparation). The six fields were up to 13 km apart from one another and soil type was for all of them
Eutric Cambisol that had developed on base moraine (Table 1). The aim of the study was to compare the influence of two
different cover cropping regimes on short-term soil organic matter cycling. Each field had 39 differential GPS (dGPS)
referenced sampling points in an unaligned sampling design. At each dGPS referenced point, soil was sampled three to four
times in three depths (0-5, 5-10 and 10-20 cm) during one long cover cropping period (August 2019 to May 2020). Fields A,
85   B, C and D had four sampling times resulting in 468 samples per field. Fields E and F had three sampling times resulting in
351 samples per field. All samples were dried at 40° C to constant weight (around 72 h) and then gently crushed and sieved to
2 mm. For the total sample size of 2574 samples, soil properties were estimated using vis–NIR soil spectroscopy, whereas 386
samples were analyzed conventionally by wet chemistry for subsequent calibration modeling. These 386 samples for laboratory
analysis were selected for each field separately using the Kennard-Stones algorithm (Kennard and Stone, 1969) to ensure a
90   cover of the whole spectral variability. Thereby, the Kennard-Stones algorithm was run with two to seven principal components
and the number of principal components was chosen that covered at least 99 % of the spectral variance and provided a reference
sample selection that well represented the different sampling times, soil depths and spatial distribution. The laboratory analysis
comprised soil organic C (SOC), total C, total N, permanganate oxidizable C (POXC) also called active C and pH.

2.2 Chemical soil analyses and its accuracy

95 Total C and N concentrations were measured on a ground aliquot by dry combustion (vario MICRO tube, Elementar, Germany). Inorganic C was analyzed for each sample in triplicates through the dissolution of carbonate in a Scheibler-apparatus with 10 % HCl solution and the measurement of the evolved CO_2 volume. SOC was then calculated as the difference between total C and the mean of the three measurements for inorganic C. POXC was measured according to the Protocol of Weil et al. (2003) with the adaption of Lucas and Weil (2012). In brief, 2.0 mL of 0.2 M $KMnO_4$ were added to 2.5 g of soil

100 and after a reaction time of 10 minutes, the absorption of the liquid was measured at 550 nm with a Spectrophotometer (UV-1800, Shimadzu Corporation, Japan). Measurement of pH was done in a 0.01 M $CaCl_2$ solution.

To estimate the lab measurement error, we took three samples per field (in total 18) where we conducted the measurements for total C, total N, POXC and pH in triplicates to calculate a standard deviation. We estimated the lab measurement error for SOC (σ_{SOC}) according to Equation 1:

105
$$\sigma_{SOC} = \sqrt{\sigma_{Total\ C}^2 + \sigma_{Inorganic\ C}^2} \tag{1}$$

Where $\sigma_{Total\ C}$ is the standard deviation of the total C measurement and $\sigma_{Inorganic\ C}$ the standard error of the inorganic C measurement because inorganic C measurement were for all samples done in triplicates. The measurement error of all 18 triplicates were then averaged to obtain the overall lab measurement error for a soil property.

To characterize the spatial variability of soil texture in the field, we measured grain size on 20 samples per field (every second

110 sampling point in 10-20 cm soil depth). Organic matter in the samples was oxidized with hydrogen peroxide (H2O2) and then grain size was measured with laser-diffraction analysis (LDA) after dispersion of the sample (22 mM socium carbonate and 18 mM sodium hexaphosphate) using a Mastersizer 2000 (Malvern Panalytical, UK). Since the LDA is underestimating the clay content compared to the standard grain size methods (Taubner et al., 2009), we measured one composite sample per field with the improved integral suspension pressure method (ISP+; Durner and Iden (2021)) on a PARIO Plus Soil Particle Analyzer

115 (METER Group, Germany/USA). We rescaled the mean sand, silt and clay content of the LDA-data to the mean of the IPS+ method, while keeping the coefficient of variation constant (See Table S3 in the supplementary material).

2.3 Spectral measurement and pre-processing of spectra

All samples were measured with a vis–NIR spectrometer (ASD FieldSpec 4 Hi-Res, Malvern Panalytical, USA) with a sampling interval of 1.4 nm from 350-1000 nm and 1.1 nm from 1000-2500 nm. The device then provides a reflectance

120 spectrum with a resolution of 1 nm and 2151 wavelengths. Measurements were done with a contact probe, containing an internal halogen bulb, which was in a fixed position and soil samples, placed in a petri dish of 1.5 cm height and 3 cm diameter, were lifted with a laboratory scissor jack until close contact with the probe to ensure a stable measurement position. For each sample, five petri dishes were filled to provide five replicate spectra per sample. Each of these five replicates consists of 30 internal repetitive scans that were automatically averaged by the device internal RS3 software. Between samples, the contact

125 probe was carefully cleaned with water and ethanol. After the 5 replicates of a sample, the calibration of the spectrometer was

checked with a 100 % reflectance white reference panel (Spectralon, 12×12 cm, Labsphere, USA). The infrared data of each sample was kept in two versions, once as reflectance spectra, as provided by the spectrometer, and once as absorbance spectra using the log(1/reflectance) transformation. Several pre-processing options and their combination were tested on both, the reflectance and the absorbance spectra: a) resampling of the spectra in an interval from 1 to 6 nm, b) cutting of the beginning

130 (350-400 nm) or the end (2450-2500 nm) of the spectra c) first or second order derivative d) Savitzky-Golay (SG) smoothing in a third order polynomial with window sizes ranging from 5 to 51, e) gap segment derivative (GSD) with window width between 5 and 51 and segment size between 1 and 21, f) standard normal variate (SNV) combined with GSD, g) SG smoothing combined with multiplicative scatter correction (MSC). All applied pre-processing techniques are frequently used in soil spectroscopy and well described in Ellinger et al. (2019). The pre-processing techniques from a) to g) led to around 100

135 meaningful combinations that were tested in model building and the final pre-processing option was selected based on the smallest RMSE.

2.4 Development and evaluation of field-specific local models

We used for all 30 local models (6 fields x 5 properties) a PLSR modelling approach (Wold et al., 1983). Model performance was assessed using the statistics of the hold-out folds of each five times repeated five-fold cross-validation because it was

140 evaluated as a robust method for smaller datasets (Kuhn and Johnson, 2013; Molinaro et al., 2005). To avoid model overfitting, we set the maximum of latent variables in the PLSR model to 12. For each number of latent variables (1, 2, …., 12) the dataset was five times randomly split into five folds of which four were used for model training and the remaining fold was held out and used for model validation. The RMSE (Eq. 2) of the hold-out samples was averaged among the five repeats resulting in a cross-validated RMSE per number of latent variables. The final number of latent variables was then chosen according to the

145 "one standard error rule" which means that instead of directly choosing the number of latent variables with smallest mean RMSE, the most parsimonious (less latent variables) model within one standard error of the mean RMSE of the optimal model was selected (Hastie et al., 2017). The "one standard error rule" was also applied during optimization of pre-processing to avoid model overfitting. The final model was trained using all training data with optimized number of latent variables.

A proper validation of a spectral model is very crucial and particularly important in this study where soil was repeatedly

150 sampled in different depths at the same GPS point. To analyze the correlation among the samples and define a grouping factor for the cross-validation, we calculated the mean Euclidean distances between all samples and compared it with the mean distance 1) between samples at the same GPS point but different depths, 2) between samples at the same point and depth but different sampling times and 3) between samples at the same point but different depth and sampling times (Fig. S1 in the supplementary material). Thereby, we observed that the soil samples from the three different soil depths sampled at the same

155 GPS-point at the same sampling time had a substantially lower mean Euclidean distance compared to the overall mean. Consequently, we grouped the samples from the same GPS point at the same sampling time and kept them in the same fold to avoid a too optimistic model evaluation during cross-validation.

Since we used a cross-validation approach on the field scale, all models showed a very small bias (see Table 2). We therefore do not discuss the bias in this paper and focus on R^2, RMSE and RPD (Eq. 3) for the evaluation and comparison of different

160 models. RMSE was calculated according to Equation 2 where $\hat{y}_i$ is the prediction of the spectral model for sample i and y_i the actual measured value for the same sample in the laboratory.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ , \tag{2}$$

RPD compares the RMSE with the standard deviation (SD, Eq. 3) of the data:

$$RPD = \frac{SD}{RMSE}\ , \tag{3}$$

165 For all model performance parameters (R^2, RMSE and RPD) of the cross-validation, we calculated the uncertainty with the standard deviation of the prediction of the hold-out folds across the five repetitions.

To classify the model performance, we combined the RPD based classification of Chang et al. (2001) and Zhang et al. (2018). We considered spectral models with RPD < 1.4 as poor, models with RPD between 1.4 and 2 as approximate, models with RPD between 2 and 3 as accurate and models with RPD > 3 as excellent. Even though in spectroscopy project of a local extent

170 the RMSE is the most important model performance parameter, RPD is the best parameter to compare models of different scales. Model metrics (R^2, RMSE and RPD) mentioned in the text are based on the cross-validation and metrics for the model calibration in Table 2 are specifically labelled as R^2_{cal}, $RMSE_{cal}$, RPD_{cal}.

2.5 Development and evaluation of general models

In addition to the field-specific local models, we built general models for the five soil properties that included all reference

175 samples (n = 386) of the six fields. Even though for this sample size an independent test-set would be more suitable than a cross-validation approach, we evaluated the model performance using the hold-out samples in the five times repeated 10-fold cross-validation, keeping, as for the local models, samples from same GPS point and same sampling time in the same fold. The first reason for not using an independent validation set is that the modelling approach of the general model should be similar to the one of the local models to make them comparable. The second reason is that a representative split of the dataset

180 into a calibration and a validation set according to the spectral variability would not result in equal number of samples per field in the validation set. Conversely, if we selected an equal sample size per field for the validation set, we would not have been able to cover the entire spectral variability. Evaluating the general models with hold-out samples of the cross-validation allowed us to calculate not only the RMSE over all samples but also the RMSE for the samples of each field individually. These field specific RMSE of the general model could then be compared with the RMSE of the local models. Since the only

185 purpose of the general models was to increase modelling efficiency for a specific combined dataset, we did not group the samples according to fields during cross-validation because the same share of samples from the same field would also be in

the prediction dataset. For the general models we cannot indicate uncertainties on a field-specific level since the folds did not always contain the same number of samples per field.

2.6 Model interpretation

190 To interpret spectral models, it is crucial to find relevant spectral features that are consistently important for a certain soil property. To identify the most important wavelength ranges in the final chosen models, we used the variable importance in projection (VIP) method first published by Wold et al. (1993) and evaluated by Chong and Jun (2005). The VIP method can deal with multicollinearity and is therefore suitable for the interpretation of spectral models as it was for example applied by Baumann et al. (2021). Wavelengths that have an above-average impact on the model have a VIP score above 1. We classified
195 spectral ranges in groups of VIP scores between 1 and 1.5, 1.5 and 2 as well as VIP scores above 2.

2.7 Assessment of site characteristics influencing model performance

To understand the reasons for the varying performance of the 35 developed spectral models, we studied the influence of various site characteristics on the models. To do so, we correlated the model performance parameters (R^2, RPD and RMSE) with field size, soil texture, carbonate content and with the correlation coefficients between SOC and total N in the dataset. With six local
200 datasets as independent variables it is hardly possible to apply statistical tests that could potentially reject a null hypothesis. Therefore, we relied on the interpretation of graphs and Pearson's moment correlation coefficients between soil properties and RMSE. Since the RMSE are estimates with uncertainties (standard deviations, see section 2.4), we used a Monte Carlo simulation and reported mean and standard deviation of the correlation coefficients after 1000 iterations. For the identified site characteristics that showed strongest trends on model performance (carbonate content, correlation coefficient between SOC
205 and N and variability in clay content) we looked for possible explanations in the spectral features. Thereby, we relied on the VIP-analysis of the trained models, on the correlation coefficients between soil properties with spectral variables and on the correlation matrices between target variables.

2.8 Data organization

All analysis were performed in R version 4.0.3 (R Core Team, 2020). The spectral datasets were analyzed using the R-package
210 *simplerspec* version 0.2.0 (Baumann, 2019) in combination with the packages *prospectr* version 0.2.1 (Stevens and Ramirez-Lopez, 2020) and *caret* version 6.0-86 (Kuhn, 2020).

3 Results

3.1 Description of the datasets

A comparison of the data distribution between the six different fields can be seen in Fig. 1 and the corresponding statistics in Table S1 in the supplementary material. The means for SOC, total N and POXC differed between the six fields, but the distribution was relatively similar for these three soil properties. The density functions for total C and pH were highly influenced by the spatial distribution of carbonate in the soil: Fields B, D and E contain samples with and without carbonate resulting in a broad distribution for both, total C and pH. All soil samples of fields A and C contained carbonate in varying concentrations resulting in a broad distribution for total C but narrow distribution for pH. Field F showed high and only slightly varying carbonate content and therefore a very narrow distribution for total C and pH. Field C had highest mean clay content and field A highest mean sand content whereas field F showed highest variability in soil texture.

3.2 Performance of spectral models

Based on RPD, 13 out of 30 local models showed an excellent performance (RPD > 3), 11 models an accurate performance (RPD > 2), five models an approximate (RPD > 1.4) and one model a poor performance (RPD > 1.4; Table 2). The six models without accurate performance were SOC, POXC and pH on field A and F.

However, the RMSE of the local models for pH of fields A (0.08 ± 0.02; mean ± standard deviation) and F (0.04 ± 0.01) were similar or smaller than the RMSE of the other three local models (between 0.08 ± 0.02 and 0.19 ± 0.03) whose performance was classified as accurate. Differently, the local models for SOC on fields A and F with only approximate performance showed a higher RMSE (2.43 ± 0.55 and 2.00 ± 0.38 g kg^{-1}) than the other accurately performing local models for SOC (between 1.07 ± 0.19 and 1.59 ± 0.28 g kg^{-1}). The five general models showed all an accurate to excellent performance with RPD ranging from 2.60 ± 0.43 to 4.16 ± 0.47.

3.3 Influence of pre-processing on spectral variability

For all 35 models, pre-processing improved the models compared to the raw spectra (see an example of pre-processing optimization for total C in Table S2 in the supplementary material). Although pre-processing was necessary for all models, we highlight that several pre-processing options performed similarly well within one standard deviation, and the differences in RMSE were often relatively small (see Table S2 in the supplementary material). Figure S2 in the supplementary material gives an overview of the best performing pre-processing techniques. Most times, the first or second order derivatives improved the models substantially. Most models performed best when the spectra were reduced to every third wavelength and models based on absorbance were a bit more frequently used than models based on reflectance. The combined application of SG filter and MSC was the most successful pre-processing while a single SG filter, GSD and SNV in combination with GSD were of minor importance. Cutting of the beginning (350-400) or end of the spectra (2450-2500) sometimes improved the model performance

but since most pre-processing steps reduce the beginning and end of the spectra, it was not possible to evaluate the cutting. Similarly, it was not possible to evaluate the window width chosen in the SG-filter because there is an interference with the resampling interval. A detailed list of the selected pre-processing options of the final models and the corresponding metrics for model performance can be found in Table 2.

The sensitivity of model performance to pre-processing can be visualized with the biplots of principal component analysis (PCA). Figure 2 shows the first three biplots of the raw spectra and the spectra that were pre-processed according to the general models of the five soil properties. The raw spectra had a very high share of the explained variance (96.8%) on the first principal component but hardly any groups according to fields could be observed with the first two principal components. All pre-processing options used for the general models decreased the explained variance on the first principal component (32.5 to 39.6 %) and a grouping according to fields could already be seen in the biplot of the first two principal components. Thereby, especially field F with highest carbonate content and field C with highest clay content often showed clear groups. Nevertheless, in the pre-processing for pH, field E with the highest pH variability shows a clear group in the first biplot and the pH variability is well represented with the first PC.

3.4 Comparison of general models with local models and lab measurement error

The overall cross-validated model metrics of the general model (black filled circle in Fig. 3) indicated over all fields for all soil properties a good performance, but the field-specific model evaluation showed distinct differences among fields. The field-specific R^2 of the general models of fields B, C, D and E was similar to the R^2 of the local model for SOC, total C, total N and POXC (only a slight slope in Fig. 3). For pH, only field C, D and E showed similar R^2 in the local and general model while fields A, B and F showed clearly higher R^2 in the local model. On the other hand, field F had clearly lower R^2 in the general model than in the local model for all soil properties except POXC. For field A, R^2 was similar between the local and the general model for SOC, total C and POXC but clearly lower for total N and pH in the general model.

The field-specific RPD of the general model was across all soil properties on average 31 % lower compared to the local models (Fig. 3). All property-field combinations of fields B, C, D and F showed at least an approximate (RPD > 1.4) performance in the general models, whereas the seven poorly (RPD < 1.4) performing property field combinations were all from fields A and F. It can therefore be concluded that the general models could not improve the low performing local models.

Field-specific RMSE of the general models was on average 47 % higher compared to the local models. However, there were substantial differences between the different fields. For Field F, the field-specific RMSE in the general models for SOC, total C, total N and pH (2.58 g kg^{-1}, 0.17 g kg^{-1} and 0.09,) were much higher compared to the local model (2.00 ± 0.38 g kg^{-1}, 0.09 ± 0.02 g kg^{-1} and 0.04 ± 0.01, respectively; Fig. 3). Similarly, for total N and pH, field A had much higher RMSE in the general model 0.22 g kg^{-1} and 0.14) than in the local model (0.14 ± 0.03 and 0.08 ± 0.02). On the other hand, fields C and E showed quite RMSE in the local and in the general model for all soil properties except total C.

The RMSE of the best local models were close to the overall lab measurement errors for SOC, total C and total N, a bit higher for pH and substantially higher for POXC (Fig. 3). The RMSE of SOC on field B (1.26 ± 0.36 g kg^{-1}) and D (1.07 ± 0.19 g kg^{-}

[1]) were within the standard deviation of the lab measurement error for SOC (1.01 ± 0.40 g kg^{-1}). The overall lab measurement error for SOC was calculated from the measurement error for total C and inorganic C and therefore for fields B and D with only little inorganic C, the lab measurement error for total C (0.83 ± 0.25 g kg^{-1}) might be the better reference. However, the

280 RMSE of the local spectral models of all fields exceeded the overall lab measurement errors between factor 1.1 and 2.4 for SOC, 1.6 and 3.2 for total C, 1.3 and 2.0 for total N, 2.3 and 4.3 for POXC and between 3.4 and 17.8 for pH. The field-specific RMSE of the general model exceeded the overall lab measurement error between factor 1.3 and 2.3 for SOC, 2.2 and 5.2 for total C, 1.5 and 3.2 for total N, 2.8 and 4.6 for POXC and between 8.3 and 19.9 for pH.

The VIP scores (Fig. 4) show that the most important wavelengths were dataset specific. It can be seen that on field B and to

285 a lower extent on field F, the same wavelengths were important in all soil properties related to soil organic matter (SOC, total C, total N and POXC), whereas on the other fields the VIP pattern of the different properties were more distinct from each other. However, for all the analyzed soil properties the wavelength ranges between 400 and 750 nm (visible) as well as 1800 and 2450 nm were most important while the range in between was of lower importance. Nevertheless, some models had VIP scores above 2 in the range between 750 and 1800 nm.

290 Prediction performance in terms of RMSE and RPD of total C of fields E and F was particularly lower in the general model than in the local model (Fig. 3). This finding can be explained with the VIP analysis (Fig. 4) that showed for the general model the most important wavelength range between 2150 and 2450 nm but for the local models of fields E and F in the range of 500 to 1020 nm. The local model for total N of field F showed very high VIP scores (>2) in a small specific range between 2345 and 2369 nm but these wavelengths were not important in the general model for total N (Fig. 4), which resulted in a much

295 lower prediction accuracy of total N for field F in the general model compared to the local model.

3.5 Site characteristics influencing model performance

We found an order of model performance with respect to R^2 and RPD in dependence of mean carbonate content, correlation coefficient between SOC and total N as well as coefficient of variation in clay content (Fig. 5). Fields A and F which showed lower model performance in terms of RPD had higher carbonate content, lower correlation coefficient between SOC and total

300 N and higher variability in soil texture (compare also with density plots in Fig. 1). However, in absolute prediction performance (RMSE) we only found for SOC and pH substantial correlations ($|r| \geq 0.46$) between RMSE and field characteristics (Fig. 6). Compared to the three mentioned field characteristics mentioned above, we found weaker influence of field size, absolute contents of sand, silt and clay or variability of carbonate content on model performance (see Fig. S3 in the supplementary material).

305 The influence of carbonate content on model performance of SOC is illustrated by plotting at each wavelength the correlation coefficients between pre-processed spectral variables and inorganic C as well as SOC content (Fig. 7). The correlation between SOC and spectral variables was higher on fields B, D and E than on fields A, C and F, which also explains the better model performance. On field A, SOC and carbonate content show a very similar correlation with spectral variables across the whole vis–NIR range, which makes it difficult to distinguish organic and inorganic C on field A resulting in an excellent performance

of total C but much lower performance for SOC (see Table 2). Even though the correlation between spectral variables and SOC content on field C was lower than on other fields (B, D and E), the very different correlation pattern of carbonate content still resulted in good model performance for SOC. Especially the ranges between 600 and 1200 nm and the peaks at 1680 nm and 2240 nm showed different spectral features for SOC and carbonate which corresponds to the high VIP scores at that wavelengths for the SOC model on field C. On field F, correlations for both carbonate content and SOC were relatively weak, whereby carbonate content showed stronger correlations with spectral variables which probably masked the spectral features of SOC resulting, as for field A, in a better model for total C than SOC.

The better model performance on field B, D and E compared to fields A, C and F also coincided with higher correlation between SOC and total N (Fig. 5). In general, correlation coefficient between target variables tended to be higher on field B, D and E compared to fields A, C and F (see Fig. 8 as example and all correlation matrices in Fig. S4 in the supplementary material).

4 Discussion

4.1 Performance of local spectral models

Most of the developed local models showed an accurate performance and confirm the suitability of vis–NIR spectroscopy in projects of local or single plot extent. The performance (based on RPD) of the two models for pH on field A and F, which were classified as only approximate or even poor, respectively, can be explained by the low variability of pH in these datasets (see Fig. 1) and is supported by the fact that these two models had the smallest RMSE values for pH (Fig. 3). This explanation does not hold for the other three local models that were also classified as only approximate because SOC and POXC on field A as well as SOC on field F showed a similar variability as on the other fields (Fig. 1), but higher RMSE. However, considering the mean SOC concentration on fields A (22.4 ± 3.7 g kg^{-1}) and F (28.6 ± 2.7 g kg^{-1}) as well as the lab measurement error (1.00 ± 0.04 g kg^{-1}), we argue that the RMSE on fields A (2.43 ± 0.55 g kg^{-1}) and F (2.00 ± 0.38 g kg^{-1}) are probably for many research projects still acceptable, especially when taking into account that a higher sample size can be analyzed for the same costs.

In agreement with literature (Soriano-Disla et al., 2014), primary properties with a direct impact in the vis–NIR range like SOC, total C, total N and POXC showed a RMSE that was closer to the lab measurement error. On the other hand, pH has only an indirect impact on the spectra and thus showed a much higher RMSE compared to the lab measurement error but the RMSE for pH in the local models (between 0.04 ± 0.01 and 0.19 ± 0.03) is probably small enough for most research purposes.

4.2 Comparison of general models with local models

The general models could not improve the prediction of low performing local models. This finding is especially interesting because in this study the general model was built with datasets of six fields that were spatially close to one another (maximal distance of 13 km), had the same soil type and the same parent material. However, the base moraine as a parent material can

be variable which we mainly observed in different soil texture and carbonate content but also in the high spectral variability (see PCA biplots in Fig. 2). In this sense, we confirm the conclusions of Seidel et al. (2019) and Ng et al. (2022) who suggested that the best solution is always to develop a local model if enough samples (> 30) are available. This conclusion is supported in this study by the quite distinctive pattern of VIP scores between the different models (Fig. 4). The overall picture shows that 345 the wavelengths between 2000 and 2450 nm followed by the visible range between 400 and 700 nm were most important for prediction of the investigated properties, which is in agreement with the literature (Munnaf and Mouazen, 2022; Soriano-Disla et al., 2014). Nevertheless, each local model has distinct and site-specific features that could not be attributed to specific soil characteristics while being important for the model development. The development of general models where different locations are aggregated in one dataset can save costs because the number of lab analysis per location can be reduced and less work is 350 required for model building. Depending on the research purpose and the required measurement accuracy, the development of general models can be a very suitable and cost-effective approach. Nevertheless, this study showed that some fields (A and F) can show a poor performance in general models, hence it is crucial to consider what locations or datasets are being combined.

4.3 Pre-processing

The selection of the optimal pre-processing scheme was crucial for model performance but strongly dependent on the dataset. 355 Often MSC was the best performing pre-processing option, which was confirmed in some studies (Cambule et al., 2012; Liu et al., 2019) but disproved in others (Knox et al., 2015; Riefolo et al., 2020). We therefore highly recommend considering MSC as a pre-processing option in spectral modelling but at the same time agree with Barra et al. (2021) that there is no general pre-processing solution that works for all datasets. The principal component analysis with the combined dataset of all fields (Fig. 2) illustrates this finding by the different grouping of individual field datasets due to different pre-processing. This 360 leads to the conclusion that studies that did not optimize the pre-processing scheme for every soil property separately did eventually not make full use of the spectroscopy, which has been shown by other studies as well (Alomar et al., 2021; Rodriguez-Febereiro et al., 2022; Singh et al., 2022). Nevertheless, the property-specific optimization of spectral pre-processing is a tedious process and constrains the fast and cost-effective application of vis–NIR spectroscopy, but some progress has recently been made by Mishra et al. (2022).

365 4.4 Site characteristics influencing model performance

We found higher model performance on fields with low carbonate content, high correlations between soil properties and low variability in clay content. We want to discuss how these identified important field characteristics influence or mask spectral features.

4.4.1 Mean carbonate content

370 We found an influence of carbonate content with lowest performance of local spectral models on fields A and F. Similar observations were made by Amare et al. (2013) and Mccarty et al. (2002) who argued that the absorbance bands of carbonate

mask those of SOC. Looking at the correlation between spectral variables and inorganic C respectively SOC (Fig. 7) we can confirm this finding but have to add that on the local scale the relative intensity of absorption bands for carbonate and SOC varied substantially between different datasets. In this context, Reeves (2010), who showed that the spectrum of a soil sample varied greatly with its carbonate content, considered the prediction of SOC in soils with high carbonate content as one of the open questions in vis–NIR spectroscopy research. An important point missing in this discussion is the measurement accuracy of SOC in the laboratory, which is strongly influenced by the presence of carbonate and the method used (Goidts et al., 2009). If the soil samples contain carbonate, often two measurements must be conducted, and SOC is calculated as the difference between total C and inorganic C. Especially with a high carbonate content, the measurement error for the inorganic C content can be a substantial share of the SOC content. The higher lab measurement error with higher carbonate content might be a possible explanation for the lower model performance on soils with high carbonate content for SOC but not for the other four soil properties where model performance (in terms of RPD) still tended to be lower than on fields with little carbonate content (Fig. 5). This confirms the above-mentioned observation of spectral interference between inorganic C and organic matter and is additionally substantiated by the result that most properties of fields A and F showed a poor performance in the general models (Fig. 3). It is known that carbonate has many more defined peaks and less interferences with organic matter in the MIR than in the vis–NIR (Reeves, 2010). Therefore, datasets that combine soil samples with high and low carbonate content might better be predicted with MIR spectroscopy. However, while all samples of field F have a high carbonate content, field A shows a broad range in carbonate contents whereby the mean carbonate content (7.1 ± 6.7 g kg^{-1}) is only slightly higher compared to the other fields. We therefore hypothesize that lower performance of field A compared to fields A, B, C and D might also have additional reasons than the field characteristics explored in this study and requires more research. The strong correlation between mean carbonate content and RMSE ($r = -0.68 \pm 0.10$; Fig. 6) can be explained by the very low variability in pH on fields with high carbonate content. The narrow pH ranges on these fields consequently lead to models for pH with low RMSE, but also low RPD (see Fig. 5).

4.4.2 Correlations between target variables

Reflectance measured with vis–NIR spectroscopy is a combined effect of all constituents present in the soil sample (Stenberg et al., 2010) and through processing and modeling one tries to distinguish the absorption feature of one specific soil property from the other constituents of the sample. Apart from pH, all our target variables were closely related to soil organic matter which was therefore for this study the most important soil constituent influencing the absorption features. In case of high correlations between target variables, that form part of soil organic matter, the modelling is easier because the same absorption features can be used for modelling the different properties which was the case for field B (see VIP analysis in Fig. 4). On the other hand, a low correlation between target variables makes it more difficult to relate absorption features of organic matter to specific soil properties, which probably contributed to the lower model performance of fields A, C and F compared to fields B, D and E. The literature shows that different soil properties related to soil organic matter (e.g. SOC and total N) can show different absorption features in the vis–NIR range (Chang and Laird, 2002; Kusumo et al., 2019), which is also supported in

405 our study (see VIP analysis in Fig. 4). However, we argue that prediction accuracy improves substantially if target variables related to soil organic matter are well correlated with each other, which was also hypothesized by Martin et al. (2002) in a one location field study.

4.4.3 Variability of clay content

Unlike Stenberg et al. (2010) and Heinze et al. (2013), we did not find a better model performance with increasing mean clay
410 content in the dataset which might also be explained by the relatively small range in mean clay contents between 18 % (Field F) and 38 % (Field C). However, we observed that the fields A and F with lower model performance also showed a higher variability in soil texture (see density plots in Fig. 1). We hypothesize that this observation is mainly an effect of our sampling design and the specific agricultural management and therefore not generalizable. Clay as well as soil organic matter are claimed to be modelled with high success rate with vis–NIR spectroscopy since they have strong absorption features (Da Silva-Sangoi
415 et al., 2022). Unfortunately, soil texture was measured on different samples than the reference dataset for the spectral modelling, so we cannot check for the correlation between soil texture and target variables. However, in this study the correlation may be relatively low for the following reason: We took samples in different depths (0-5, 5-10 and 10-20 cm) within the past tillage layer and therefore expect that the soil texture is homogenized across the sampling depth. Since all fields are now under organic reduced tillage management, the three soil layers show quite distinct soil organic matter content (see
420 Fig. S5 in the supplementary material) but very probably similar soil texture. Therefore, a high (horizontal) variability in soil texture on a field (e. g. clay content) without strong correlation to organic matter could have added "noise" to the spectrum which worsened the prediction accuracy in our specific sampling design. Nevertheless, in untilled soils or more distinct depth segments a high variability in soil texture may not be a disadvantage in vis–NIR modeling because it might also be correlated with organic matter.

425 ## 5 Conclusion

This study investigated the impact of site characteristics on vis–NIR modeling performances and compared a local and a general modeling approach. Among the 35 models, 29 performed accurate or even excellent whereby the RMSE was close to the lab measurement error and achieved prediction accuracies are probably for many research purposes acceptable. The local models with lowest performance, were all from field A and F and we found three field characteristics in their datasets that
430 interfered with model performance. Fields A and F had higher mean carbonate content, lower correlation between target soil properties and higher variability in soil texture compared to the other fields. The influence of soil texture variability was mainly an issue in this specific sampling design whereas the influence of carbonate content and correlation between soil properties can probably be generalized due to observed spectral features and VIP analysis. Before starting a local vis–NIR project, testing for inorganic C content can be done relatively easily but it is almost impossible to know beforehand the correlations between

different soil properties. One can only be aware of the correlation issue and consider potential gradients of soil properties while designing the sampling design which is probably more important and feasible in disturbed or agricultural soils than in natural undisturbed soils. In terms of efficiency in data collection, we conclude that in a region, several target sites (or agricultural fields) with low carbonate content can be combined in a general model with only a minor reduction in model performance. A general model for multiple target sites then also allows to reduce the number of wet chemistry analyses. Whether or not several target sites with high carbonate content can be combined in one general model using vis–NIR spectroscopy is a question that requires further research. However, since carbonates show less interferences with organic matter in the MIR than in the vis– NIR spectral range, soil samples from sites with high carbonate content might be better predicted with MIR spectroscopy. Yet, the application of laboratory vis–NIR spectroscopy in projects of local extent provides the opportunity to increase the spatial or temporal resolution in a sampling design cost-effectively with only minor decreases in measurement accuracy.

References

Allory, V., Cambou, A., Moulin, P., Schwartz, C., Cannavo, P., Vidal-Beaudet, L., and Barthes, B. G.: Quantification of soil organic carbon stock in urban soils using visible and near infrared reflectance spectroscopy (VNIRS) in situ or in laboratory conditions, Sci. Total Environ., 686, 764-773, https://doi.org/10.1016/j.scitotenv.2019.05.192, 2019.

Alomar, S., Mireei, S. A., Hemmat, A., Masoumi, A. A., and Khademi, H.: Comparison of Vis/SWNIR and NIR spectrometers combined with different multivariate techniques for estimating soil fertility parameters of calcareous topsoil in an arid climate, Biosys. Eng., 201, 50-66, https://doi.org/10.1016/j.biosystemseng.2020.11.007, 2021.

Amare, T., Hergarten, C., Hurni, H., Wolfgramm, B., Yitaferu, B., and Selassie, Y. G.: Prediction of Soil Organic Carbon for Ethiopian Highlands Using Soil Spectroscopy, ISRN Soil Sci., 2013, 720589, https://doi.org/10.1155/2013/720589, 2013.

Angelopoulou, T., Balafoutis, A., Zalidis, G., and Bochtis, D.: From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation - A Review, Sustainability-Basel, 12, https://doi.org/10.3390/su12020443, 2020.

Barra, I., Haefele, S. M., Sakrabani, R., and Kebede, F.: Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances - A review, Trac-Trend. Anal. Chem., 135, https://doi.org/10.1016/j.trac.2020.116166, 2021.

Baumann, P.: philipp-baumann/simplerspec: Beta release simplerspec 0.1.0 for zenodo, Zenodo [code], https://doi.org/10.5281/zenodo.3303637, 2019.

Baumann, P., Lee, J., Frossard, E., Schönholzer, L. P., Diby, L., Hgaza, V. K., Kiba, D. I., Sila, A., Sheperd, K., and Six, J.: Estimation of soil properties with mid-infrared soil spectroscopy across yam production landscapes in West Africa, Soil, 7, 717-731, https://doi.org/10.5194/soil-7-717-2021, 2021.

Breure, T. S., Prout, J. M., Haefele, S. M., Milne, A. E., Hannam, J. A., Moreno-Rojas, S., and Corstanje, R.: Comparing the effect of different sample conditions and spectral libraries on the prediction accuracy of soil properties from near- and mid-infrared spectra at the field-scale, Soil Till. Res., 215, https://doi.org/10.1016/j.still.2021.105196, 2022.

Brown, D. J.: Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed, Geoderma, 140, 444-453, https://doi.org/10.1016/j.geoderma.2007.04.021, 2007.

470 Camargo, L. A., do Amaral, L. R., dos Reis, A. A., Brasco, T. L., and Magalhaes, P. S. G.: Improving soil organic carbon mapping with a field-specific calibration approach through diffuse reflectance spectroscopy and machine learning algorithms, Soil Use Manage., 38, 292-303, https://doi.org/10.1111/sum.12775, 2022.

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., and Smaling, E. M. A.: Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique, Geoderma, 183, 41-48, https://doi.org/10.1016/j.geoderma.2012.03.011, 2012.

475 Chang, C. W. and Laird, D. A.: Near-infrared reflectance spectroscopic analysis of soil C and N, Soil Sci., 167, 110-116, https://doi.org/10.1097/00010694-200202000-00003, 2002.

Chang, C. W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R.: Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties, Soil Sci. Soc. Am. J., 65, 480-490, https://doi.org/10.2136/sssaj2001.652480x, 2001.

480 Chong, I. G. and Jun, C. H.: Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. Syst., 78, 103-112, https://doi.org/10.1016/j.chemolab.2004.12.011, 2005.

da Silva-Sangoi, D. V., Horst, T. Z., Moura-Bueno, J. M., Dalmolin, R. S. D., Sebem, E., Gebler, L., and Santos, M. D.: Soil organic matter and clay predictions by laboratory spectroscopy: Data spatial correlation, Geoderma Reg., 28, https://doi.org/10.1016/j.geodrs.2022.e00486, 2022.

485 Durner, W. and Iden, S. C.: The improved integral suspension pressure method (ISP plus) for precise particle size analysis of soil and sedimentary materials, Soil Till. Res., 213, 105086, https://doi.org/10.1016/j.still.2021.105086, 2021.

Ellinger, M., Merbach, I., Werban, U., and Liess, M.: Error propagation in spectrometric functions of soil organic carbon, Soil, 5, 275-288, https://doi.org/10.5194/soil-5-275-2019, 2019.

Goidts, E., Van Wesemael, B., and Crucifix, M.: Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales, Eur. J. Soil Sci., 60, 723-739, https://doi.org/doi:10.1111/j.1365-2389.2009.01157.x, 2009.

490 Greenberg, I., Seidel, M., Vohland, M., Koch, H. J., and Ludwig, B.: Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies, Geoderma, 409, https://doi.org/10.1016/j.geoderma.2021.115614, 2022.

Grunwald, S., Yu, C. R., and Xiong, X.: Transferability and Scalability of Soil Total Carbon Prediction Models in Florida, USA, Pedosphere, 28, 856-872, https://doi.org/10.1016/s1002-0160(18)60048-7, 2018.

495 Hastie, T., Tibshirani, R., and Friedman, J. H.: The elements of statistical learning : data mining, inference, and prediction, Second edition, corrected at 12th printing 2017, Springer series in statistics, Springer, New York, NY, https://doi.org/10.1007/978-0-387-84858-7, 2017.

Heinze, S., Vohland, M., Joergensen, R. G., and Ludwig, B.: Usefulness of near-infrared spectroscopy for the prediction of
500   chemical and biological soil properties in different long-term experiments, J. Plant Nutr. Soil Sci., 176, 520-528,
https://doi.org/10.1002/jpln.201200483, 2013.

Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., and Vohland, M.: In situ and laboratory soil spectroscopy with portable visible-
to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils, Geoderma, 355,
https://doi.org/10.1016/j.geoderma.2019.113900, 2019.

505   Kennard, R. W. and Stone, L. A.: Computer aided design of experiments, Technometrics, 11, 137-148,
https://doi.org/10.2307/1266770, 1969.

Knox, N. M., Grunwald, S., McDowell, M. L., Bruland, G. L., Myers, D. B., and Harris, W. G.: Modelling soil carbon fractions
with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy, Geoderma, 239, 229-239,
https://doi.org/10.1016/j.geoderma.2014.10.019, 2015.

510   Kuang, B. and Mouazen, A. M.: Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on
three European farms, Eur. J. Soil Sci., 62, 629-636, https://doi.org/10.1111/j.1365-2389.2011.01358.x, 2011.

Kuang, B. and Mouazen, A. M.: Influence of the number of samples on prediction error of visible and near infrared
spectroscopy of selected soil properties at the farm scale, Eur. J. Soil Sci., 63, 421-429, https://doi.org/10.1111/j.1365-
2389.2012.01456.x, 2012.

515   Kuhn, M.: caret: Classification and Regression Training, R package [code], https://doi.org/10.18637/jss.v028.i05, 2020.

Kuhn, M. and Johnson, K.: Applied predictive modeling, Springer, New York, https://doi.org/10.1007/978-1-4614-6849-3,
2013.

Kusumo, B. H., Sukartono, S., Bustan, B., and Purwanto, Y. A.: Total nitrogen in rice paddy field independently predicted from
soil carbon using Near Infrared Reflectance Spectroscopy (NIRS), 4th Annual Applied Science and Engineering
520   Conference (AASEC), Univ Pendidikan Indonesia, Sch Postgraduate Studies, Tech & Vocat Educ St, Bali, INDONESIA,
Apr 24, https://doi.org/10.1088/1742-6596/1402/2/022096, 2019.

Li, H. Y., Jia, S. Y., and Le, Z. C.: Prediction of Soil Organic Carbon in a New Target Area by Near-Infrared Spectroscopy:
Comparison of the Effects of Spiking in Different Scale Soil Spectral Libraries, Sensors, 20,
https://doi.org/10.3390/s20164357, 2020.

525   Liu, S., Shen, H., Chen, S., Zhao, X., Biswas, A., Xiaolin, J., Shi, Z., and Fang, J.: Estimating forest soil organic carbon content
using vis-NIR spectroscopy: Implications for large-scale soil carbon spectroscopic assessment, Geoderma, 348, 37-44,
https://doi.org/10.1016/j.geoderma.2019.04.003, 2019.

Lobsey, C. R., Viscarra Rossel, R. A., Roudier, P., and Hedley, C. B.: rs-local data-mines information from spectral libraries to
improve local calibrations, Eur. J. Soil Sci., 68, 840-852, https://doi.org/10.1111/ejss.12490, 2017.

530   Lucas, S. T. and Weil, R. R.: Can a Labile Carbon Test be Used to Predict Crop Responses to Improve Soil Organic Matter
Management?, Agron. J., 104, 1160-1170, https://doi.org/10.2134/agronj2011.0415, 2012.

Martin, P. D., Malley, D. F., Manning, G., and Fuller, L.: Determination of soil organic carbon and nitrogen at the field level using near-infrared spectroscopy, Can. J. Soil Sci., 82, 413-422, https://doi.org/10.4141/s01-054, 2002.

McCarty, G., Reeves, J., Reeves, V., Follett, R., and Kimble, J.: Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement, Soil Sci. Soc. Am. J., 66, 640-646, https://doi.org/10.2136/sssaj2002.6400a, 2002.

Mishra, P., Roger, J. M., Marini, F., Biancolillo, A., and Rutledge, D. N.: Pre-processing ensembles with response oriented sequential alternation calibration (PROSAC): A step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling, Chemom. Intell. Lab. Syst., 222, https://doi.org/10.1016/j.chemolab.2022.104497, 2022.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M.: Prediction error estimation: a comparison of resampling methods, Bioinformatics, 21, 3301-3307, https://doi.org/10.1093/bioinformatics/bti499, 2005.

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

Ng, W., Minasny, B., Jones, E., and McBratney, A.: To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library, Geoderma, 406, https://doi.org/10.1016/j.geoderma.2021.115501, 2022.

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing [code], https://www.R-project.org , 2020.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Dematte, J. A. M., and Scholten, T.: The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets, Geoderma, 195, 268-279, https://doi.org/10.1016/j.geoderma.2012.12.014, 2013.

Reeves, J. B.: Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done?, Geoderma, 158, 3-14, https://doi.org/10.1016/j.geoderma.2009.04.005, 2010.

Riefolo, C., Castrignano, A., Colombo, C., Conforti, M., Ruggieri, S., Vitti, C., and Buttafuoco, G.: Investigation of soil surface organic and inorganic carbon contents in a low-intensity farming system using laboratory visible and near-infrared spectroscopy, Arch. Agron. Soil Sci., 66, 1436-1448, https://doi.org/10.1080/03650340.2019.1674446, 2020.

Rodriguez-Febereiro, M., Dafonte, J., Fandino, M., Cancela, J. J., and Rodriguez-Perez, J. R.: Evaluation of Spectroscopy and Methodological Pre-Treatments to Estimate Soil Nutrients in the Vineyard, Remote Sens., 14, https://doi.org/10.3390/rs14061326, 2022.

Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S., and Vohland, M.: Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations, Geoderma, 354, https://doi.org/10.1016/j.geoderma.2019.07.014, 2019.

565 Shen, Z. F., Ramirez-Lopez, L., Behrens, T., Cui, L., Zhang, M. X., Walden, L., Wetterlind, J., Shi, Z., Sudduth, K. A., Song, Y. Z., Catambay, K., and Rossel, R. A. V.: Deep transfer learning of global spectra for local soil carbon monitoring, Isprs Journal of Photogrammetry and Remote Sensing, 188, 190-200, https://doi.org/10.1016/j.isprsjprs.2022.04.009, 2022.

Singh, K., Aitkenhead, M., Fidelis, C., Yinil, D., Sanderson, T., Snoeck, D., and Field, D. J.: Optimization of spectral pre-processing for estimating soil condition on small farms, Soil Use Manage., 38, 150-163, https://doi.org/10.1111/sum.12684, 2022.

570 Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., and McLaughlin, M. J.: The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties, Appl. spectrosc. Rev., 49, 139-186, https://doi.org/10.1080/05704928.2013.811081, 2014.

Stenberg, B., Rossel, R. A. V., Mouazen, A. M., and Wetterlind, J.: Visible and near infrared spectroscopy in soil science, in: Adv. Agron., edited by: Sparks, D. L., Advances in Agronomy, 163-215, https://doi.org/10.1016/s0065-2113(10)07005-7, 2010.

Stevens, A. S. and Ramirez-Lopez, L.: An introduction to the prospectr package, R package [code], https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr.html, 2020.

Taubner, H., Roth, B., and Tippkotter, R.: Determination of soil texture: Comparison of the sedimentation method and the laser-diffraction analysis, J. Plant Nutr. Soil Sci., 172, 161-171, https://doi.org/10.1002/jpln.200800085, 2009.

580 Weil, R. R., Islam, K. R., Stine, M. A., Gruver, J. B., and Samson-Liebig, S. E.: Estimating active carbon for soil quality assessment: A simplified method for laboratory and field use, Am. J. Altern. Agric., 18, 3-17, https://www.jstor.org/stable/pdf/44503242.pdf, 2003.

Wetterlind, J. and Stenberg, B.: Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples, Eur. J. Soil Sci., 61, 823-843, https://doi.org/10.1111/j.1365-2389.2010.01283.x, 2010.

Wold, S., Johansson, E., and Cocchi, M: PLS-partial least squares projections to latent structures, in: 3D QSAR in drug design, edited by: Kubinyi, H., Folkers, G., and Martin, Y., Escom, Leiden, 523– 550, https://doi.org/10.1007/0-306-46858-1, 1993.

Wold, S., Martens, H., and Wold, H.: The multivariate calibration problem in chemistry solved by the PLS method, Matrix Pencils, Berlin, Heidelberg, 286-293, https://doi.org/10.1007/BFb0062108, 1983.

Zhang, L., Yang, X. M., Drury, C., Chantigny, M., Gregorich, E., Miller, J., Bittman, S., Reynolds, W. D., and Yang, J. Y.: Infrared spectroscopy estimation methods for water-dissolved carbon and amino sugars in diverse Canadian agricultural soils, Can. J. Soil Sci., 98, 484-499, https://doi.org/10.1139/cjss-2018-0027, 2018.

Zhao, D. X., Arshad, M., Wang, J., and Triantafilis, J.: Soil exchangeable cations estimation using Vis-NIR spectroscopy in different depths: Effects of multiple calibration models and spiking, Comput. Electron. Agric., 182, https://doi.org/10.1016/j.compag.2021.105990, 2021.

Table 1: Description of the datasets of the six different fields A to F. All fields were classified as Eutric Cambisol developed on base moraine. Soil texture was measured with the improved integral suspension pressure method (ISP+)

Field	Coordinates	Elevation [m above sea level]	Area [ha]	Mean soil texture (Sand/Silt/Clay) [%]	Number of samples	
					Spectroscopy	Wet chemistry
A	47° 40' 58" N / 08° 45' 54" E	420	0.84	Sandy loam (50/29/21)	468	70
B	47° 40' 54" N / 08° 46' 05" E	420	0.67	Sandy loam (44/35/20)	468	70
C	47° 38' 01" N / 08° 57' 02" E	600	0.44	Sandy loam (27/35/38)	468	70
D	47° 38' 43" N / 08° 42' 58" E	460	0.64	Clay loam (28/44/28)	468	70
E	47° 38' 49" N / 08° 43' 06" E	460	1.05	Sandy loam (30/48/23)	351	53
F	47° 34' 22" N / 08° 48' 52" E	380	0.3	Sandy loam (39/43/18)	351	53

600

Table 2: Description of applied pre-processing and model performance of the final chosen models using a Partial Least Square regression. The local models (fields A to F) were evaluated with 5 times repeated 5-fold cross-validation and the general models (All) with 5 times repeated 10-fold cross-validation. Model metrics of Cross-validation are indicated as mean with the standard deviation across the repeats in brackets. RMSE = Root mean square error, RPD = ratio of performance to deviation, Refl. = Reflectance, Abs. = Absorbance, SG = Savitzky-Golay filter (m = order of derivative, w = window width), SNV = standard normal variate, GSD = gap segment derivative (m = derivative, w = window width, s = segment size), MSC = multiplicative scatter correction

Field	Property	Range of wavelengths / interval [nm]	Pre-processing	Latent variable	n	Calibration				Cross-validation				Model performance
						R^2_{cal}	$Bias_{cal}$	$RMSE_{cal}$	RPD_{cal}	R^2	Bias	RMSE	RPD	
A	SOC [g kg^{-1}]	410-2500 / 6	Abs., SG (m = 1, w=35)	2	70	0.59	0.00	2.35	1.58	0.55 (0.14)	0.01 (0.69)	2.43 (0.55)	1.50 (0.30)	Approximate
	POXC [g kg^{-1}]	350-2500 / 2	Refl., SNV, GSD (m=2, w=31, s=1)	7	70	0.81	0.00	0.05	2.28	0.64 (0.17)	0.00 (0.02)	0.07 (0.01)	1.65 (0.41)	Approximate
	Total N [g kg^{-1}]	370-2480 / 6	Refl., SG (m=1, w=21), MSC	7	70	0.87	0.00	0.11	2.77	0.79 (0.11)	-0.01 (0.05)	0.14 (0.03)	2.18 (0.61)	Accurate
	Total C [g kg^{-1}]	390-2500 / 4	Abs., SNV, GSD (m=2, w=21, s=1)	6	70	0.94	0.00	2.14	4.21	0.88 (0.09)	0.00 (0.84)	2.63 (0.66)	3.48 (1.41)	Excellent
	pH	410-2500 / 4	Abs., SG (m = 1, w=35)	5	70	0.74	0.00	0.06	1.97	0.63 (0.21)	0.00 (0.03)	0.08 (0.02)	1.70 (0.60)	Approximate
B	SOC [g kg^{-1}]	360-2500 / 5	Abs., SG (m=2, w=21), MSC	7	70	0.98	0.00	0.66	6.47	0.91 (0.05)	-0.04 (0.37)	1.26 (0.36)	3.46 (1.12)	Excellent
	POXC [g kg^{-1}]	360-2480 / 3	Abs., SG (m=2, w=21), MSC	4	70	0.94	0.00	0.03	4.08	0.84 (0.12)	0.00 (0.02)	0.05 (0.01)	2.60 (0.74)	Accurate
	Total N [g kg^{-1}]	360-2480 / 5	Abs., SG (m=2, w=21), MSC	4	70	0.93	0.00	0.10	3.87	0.87 (0.08)	0.00 (0.04)	0.13 (0.03)	2.85 (0.92)	Accurate
	Total C [g kg^{-1}]	370-2500 / 5	Abs., SG (m=2, w=21), MSC	10	70	0.99	0.00	0.51	9.64	0.93 (0.03)	-0.05 (0.40)	1.29 (0.25)	3.65 (0.84)	Excellent
	pH	350-2500 / 3	Abs., SG (m=1, w=21), MSC	7	70	0.98	0.00	0.07	6.64	0.83 (0.07)	0.00 (0.06)	0.19 (0.03)	2.46 (0.51)	Accurate
C	SOC [g kg^{-1}]	370-2480 / 1	Abs., SNV, GSD (m=1, w=11, s=1)	7	70	0.90	0.00	1.02	3.11	0.77 (0.09)	0.03 (0.46)	1.59 (0.28)	2.05 (0.46)	Accurate
	POXC [g kg^{-1}]	370-2440 / 3	Refl., SG (m=2, w =21)	7	70	0.93	0.00	0.03	3.80	0.77 (0.15)	0.00 (0.01)	0.05 (0.01)	2.30 (0.81)	Accurate
	Total N [g kg^{-1}]	350-2460 / 4	Abs., SG (m=2, w=21), MSC	7	70	0.97	0.00	0.05	5.87	0.90 (0.06)	0.00 (0.03)	0.09 (0.02)	3.22 (0.97)	Excellent
	Total C [g kg^{-1}]	350-2500 / 3	Refl., SG (m=1, w=21), MSC	10	70	0.97	0.00	0.92	5.69	0.93 (0.03)	-0.07 (0.34)	1.44 (0.29)	3.74 (0.98)	Excellent
	pH	390-2500 / 5	Abs., SG (m=2, w=21), MSC	6	70	0.89	0.00	0.05	3.09	0.77 (0.12)	0.00 (0.03)	0.08 (0.02)	2.00 (0.59)	Accurate
D	SOC [g kg^{-1}]	390-2500 / 3	Abs., SNV, GSD (m=2, w=21, s=1)	6	70	0.97	0.00	0.81	6.01	0.95 (0.02)	-0.01 (0.35)	1.07 (0.19)	4.74 (1.23)	Excellent
	POXC [g kg^{-1}]	390-2460 / 6	Refl., SG (m=2, w=21)	7	69	0.95	0.00	0.03	4.72	0.92 (0.03)	0.00 (0.01)	0.05 (0.01)	3.47 (0.65)	Excellent
	Total N [g kg^{-1}]	370-2500 / 4	Abs., SG (m=2, w=21), MSC	6	70	0.98	0.00	0.06	7.30	0.95 (0.04)	0.01 (0.03)	0.11 (0.02)	4.66 (1.16)	Excellent

	Property	Range / order	Preprocessing		n									Rating
	Total C [g kg⁻¹]	350-2500 / 2	Refl., SNV, GSD (m=2, w=21, s=1)	6	70	0.97	0.00	1.15	5.44	0.93 (0.03)	0.02 (0.45)	1.61 (0.39)	4.07 (1.04)	Excellent
	pH	350-2500 / 6	Refl., SG (m=2, w=21), MSC	9	70	0.99	0.00	0.06	9.79	0.95 (0.02)	0.00 (0.04)	0.13 (0.03)	4.83 (1.23)	Excellent
E	SOC [g kg⁻¹]	350-2500 / 3	Abs., SG (m=1, w=25)	3	53	0.82	0.00	1.25	2.35	0.79 (0.11)	-0.05 (0.53)	1.40 (0.42)	2.20 (0.70)	Accurate
	POXC [g kg⁻¹]	350-2500 / 4	Abs., GSD (m=2, w=21, s=21)	4	53	0.82	0.00	0.05	2.41	0.82 (0.11)	0.00 (0.02)	0.05 (0.02)	2.33 (0.69)	Accurate
	Total N [g kg⁻¹]	350-2500 / 3	Abs., SG (m=2, w=21)	4	53	0.94	0.00	0.07	4.12	0.90 (0.04)	0.00 (0.03)	0.10 (0.02)	3.10 (0.57)	Excellent
	Total C [g kg⁻¹]	360-2500 / 3	Refl., SG (m=1, w=21), MSC	6	53	0.98	0.00	1.20	7.83	0.96 (0.03)	0.04 (0.56)	1.72 (0.51)	5.27 (1.85)	Excellent
	pH	350-2500 / 4	Refl., SG (m=1, w=21), MSC	7	53	0.98	0.00	0.10	7.15	0.95 (0.03)	0.01 (0.08)	0.16 (0.05)	4.57 (1.91)	Excellent
F	SOC [g kg⁻¹]	350-2500 / 3	Abs., SG (m=1, w=21), MSC	4	53	0.66	0.00	1.59	1.73	0.51 (0.18)	0.01 (0.72)	2.00 (0.38)	1.43 (0.39)	Approximate
	POXC [g kg⁻¹]	380-2500 / 2	Refl., GSD (m =2, w=21, s=21)	5	53	0.86	0.00	0.03	2.72	0.76 (0.16)	0.00 (0.01)	0.03 (0.00)	1.96 (0.60)	Approximate
	Total N [g kg⁻¹]	350-2500 / 3	Abs., SG (m=1, w=21), MSC	5	53	0.92	0.00	0.06	3.47	0.83 (0.10)	0.00 (0.04)	0.09 (0.02)	2.51 (0.84)	Accurate
	Total C [g kg⁻¹]	370-2500 / 6	Abs., SG (m=1, w=21), MSC	5	53	0.84	0.00	0.96	2.49	0.72 (0.18)	0.01 (0.56)	1.29 (0.24)	1.82 (0.54)	Accurate
	pH	370-2500 / 4	Refl., SG (m=1, w=51)	1	53	0.23	0.00	0.04	1.15	0.30 (0.20)	0.00 (0.02)	0.04 (0.01)	1.14 (0.19)	Poor
All	SOC [g kg⁻¹]	350-2500 / 3	Refl., SNV, GSD (m=2, w=5, s=1)	7	386	0.92	0.00	1.66	3.65	0.90 (0.04)	-0.02 (0.39)	1.93 (0.32)	3.21 (0.57)	Excellent
	POXC [g kg⁻¹]	350-2500 / 1	Refl., SG (m=1, w=21), MSC	8	385	0.88	0.00	0.05	2.90	0.85 (0.05)	0.00 (0.01)	0.06 (0.01)	2.60 (0.43)	Accurate
	Total N [g kg⁻¹]	350-2500 / 4	Abs., SG (m=2, w=11)	7	386	0.92	0.00	0.14	3.53	0.89 (0.04)	0.00 (0.03)	0.16 (0.02)	3.06 (0.46)	Excellent
	Total C [g kg⁻¹]	350-2500 / 2	Abs., SNV, GSD (m=2, w=5, s=1)	6	386	0.96	0.00	2.31	5.04	0.94 (0.01)	0.00 (0.53)	2.79 (0.30)	4.16 (0.47)	Excellent
	pH	350-2500 / 3	Refl., SG (m=2, w=21), MSC	9	386	0.95	0.00	0.12	4.40	0.90 (0.03)	0.00 (0.02)	0.15 (0.02)	3.34 (0.59)	Excellent

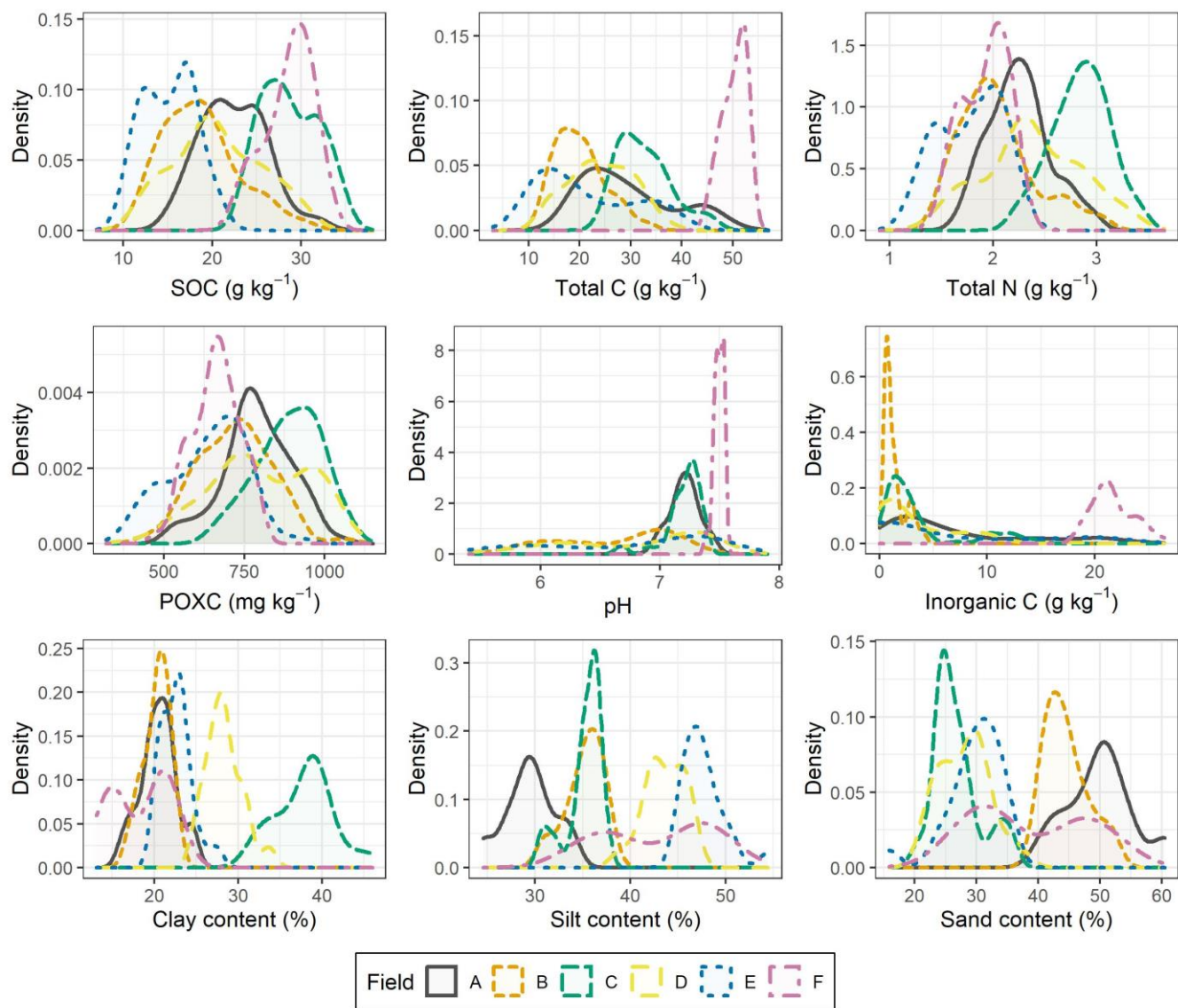



Figure 1: Density plots of the reference samples for the five target properties (SOC, total C, total N, POXC and pH) and inorganic C. Field A to D contained each 70 samples and field E and F each 53 samples. Soil texture was analyzed on 20 samples per field.

610

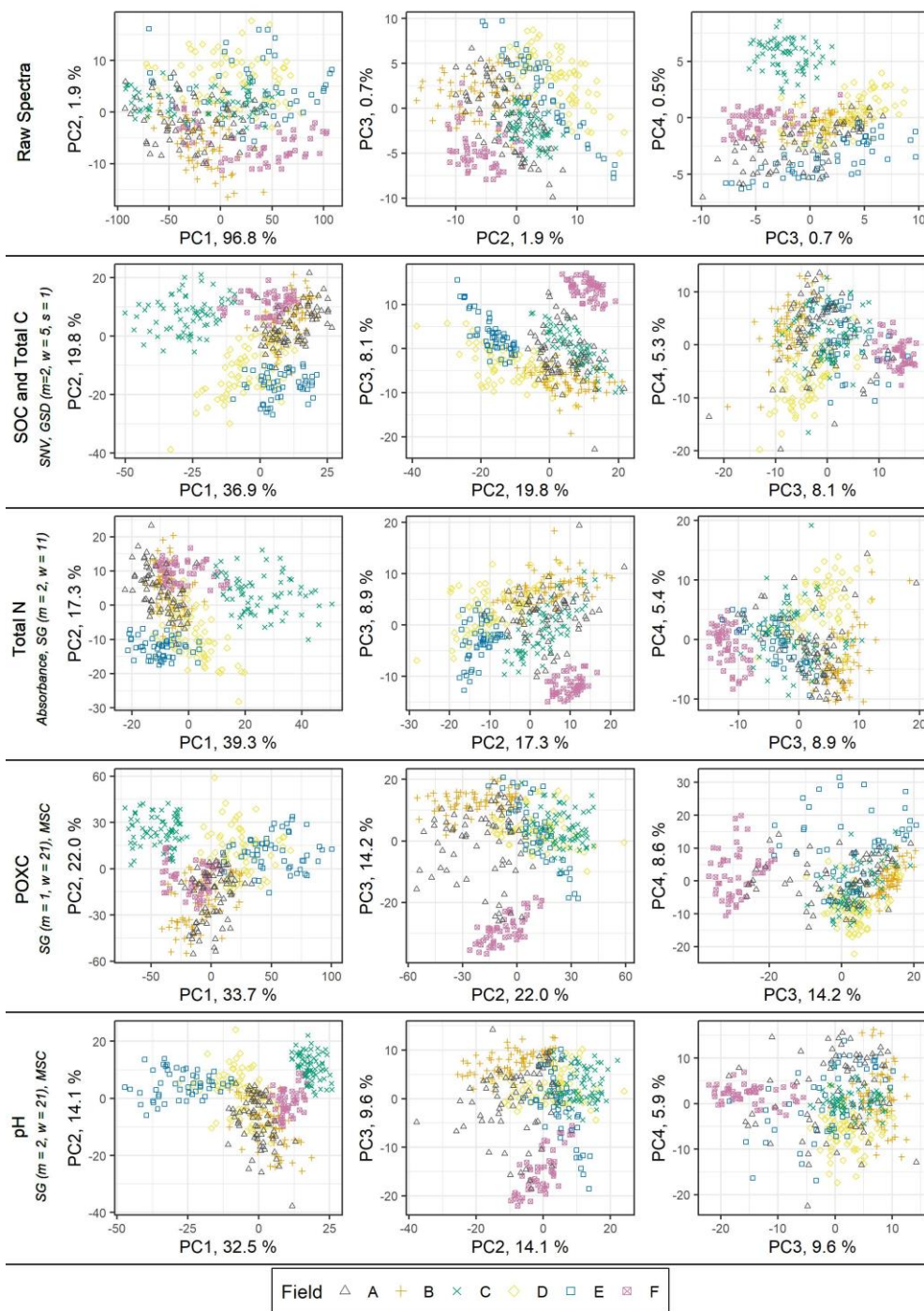
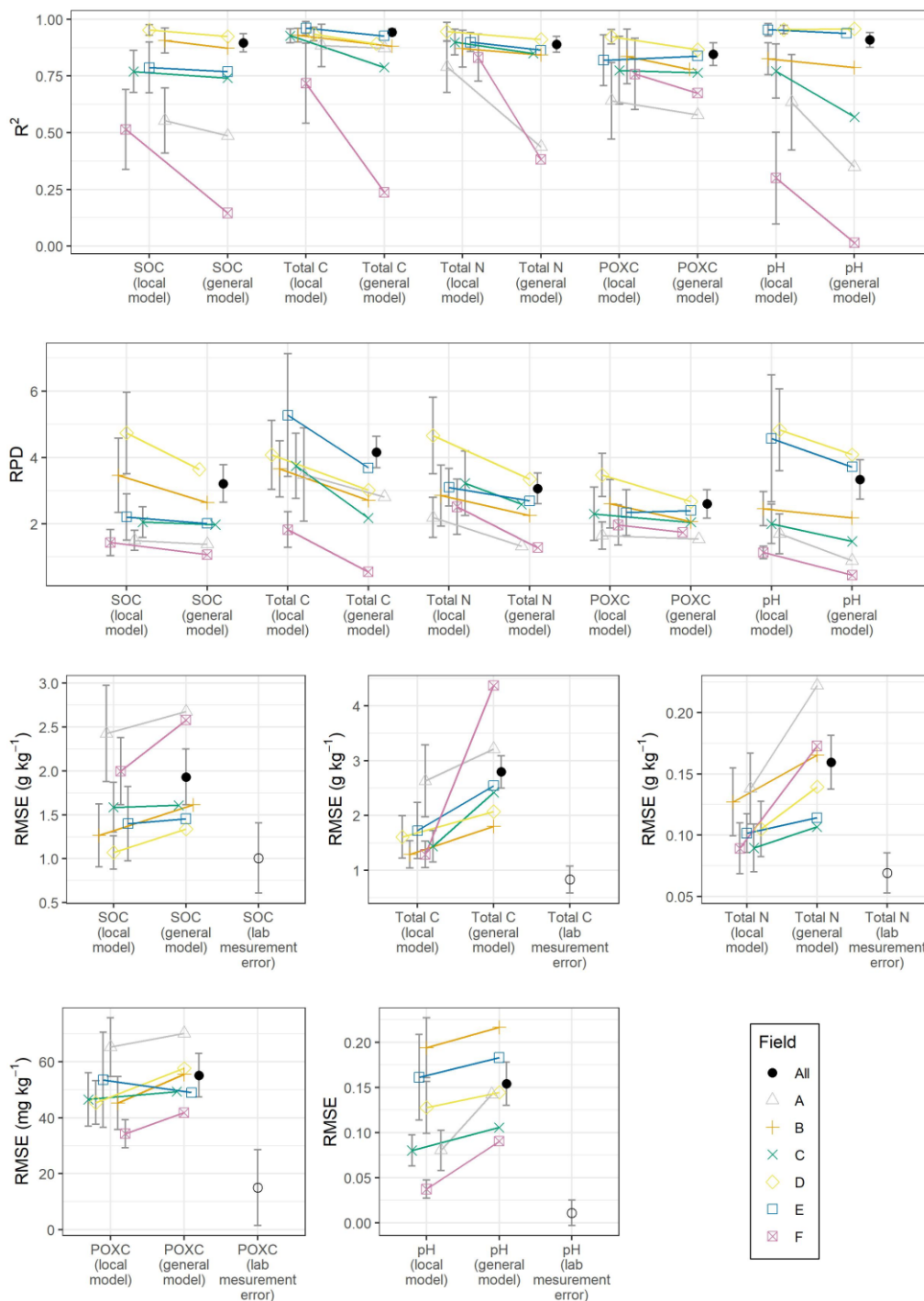



Figure 2: Biplots of principle component analysis with the first four principal components for the raw spectra and the pre-processed spectra according to the properties SOC, total C, total N, POXC and pH. The pre-processing is indicated in the figure and except for total N it was conducted on reflectance spectra (SG = Savitzky-Golay filter (m = order of derivative, w = window width), SNV = standard normal variate, GSD = gap segment derivative (m = derivative, w = window width, s = segment size), MSC = multiplicative scatter correction)

620 **Figure 3: R², ratio of performance to deviation (RPD) and root mean square error (RMSE) calculated from the local models and field-specifically calculated from the general model for the six fields (A – F) and the five soil properties (SOC, total C, total N, POXC and pH). The error bars for RMSE of spectral models represent standard deviations across the repeats in the cross-validation. The overall RMSE of the general model is indicated with a black filled circle and the label "All". The RMSE are compared with the error of the lab measurements (mean standard error of 18 triplicates indicated with standard deviation)**

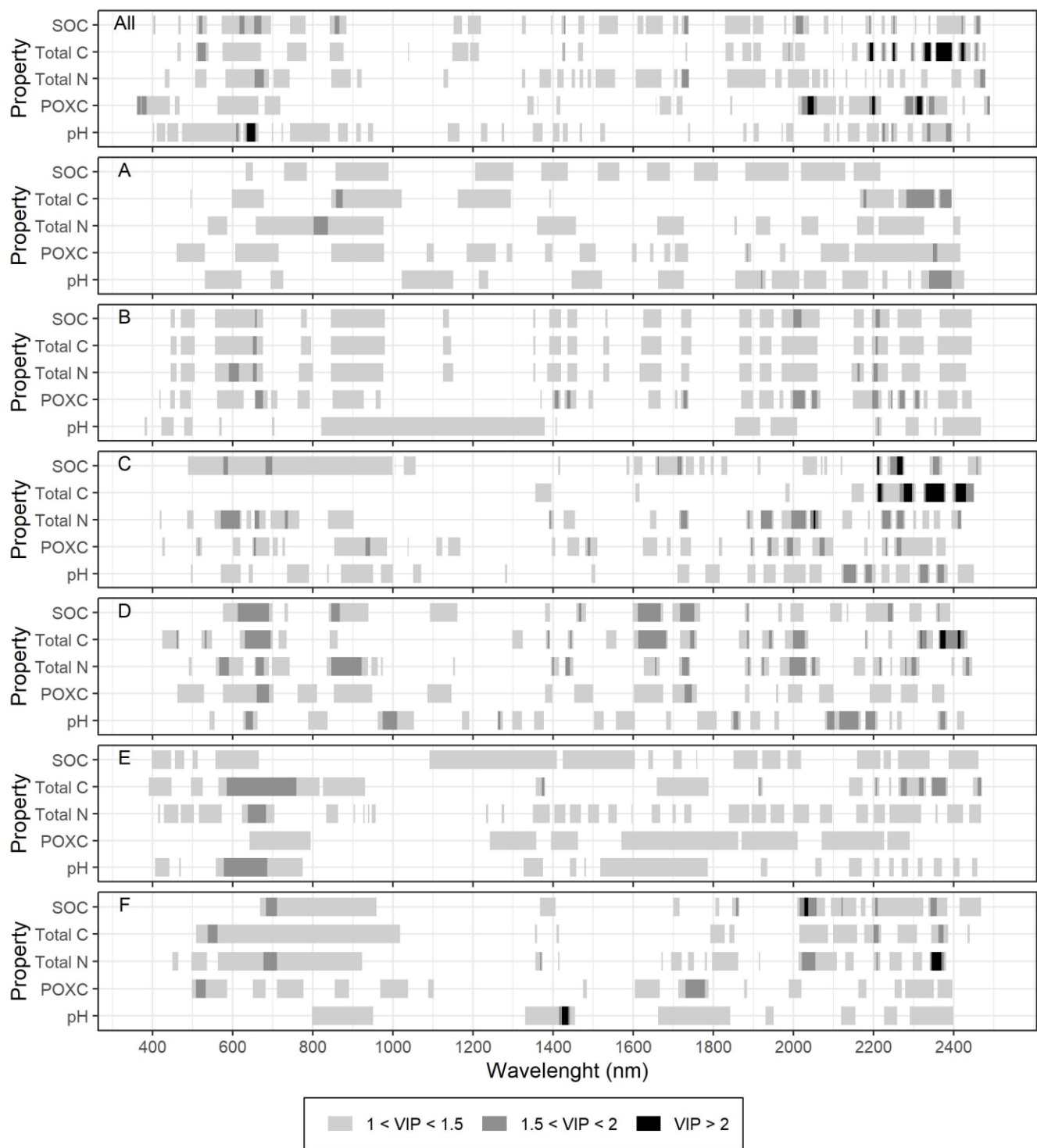



Figure 4: Variable importance in projection (VIP) for the local models of fields A to F and the general model that combined the datasets of all fields (All).

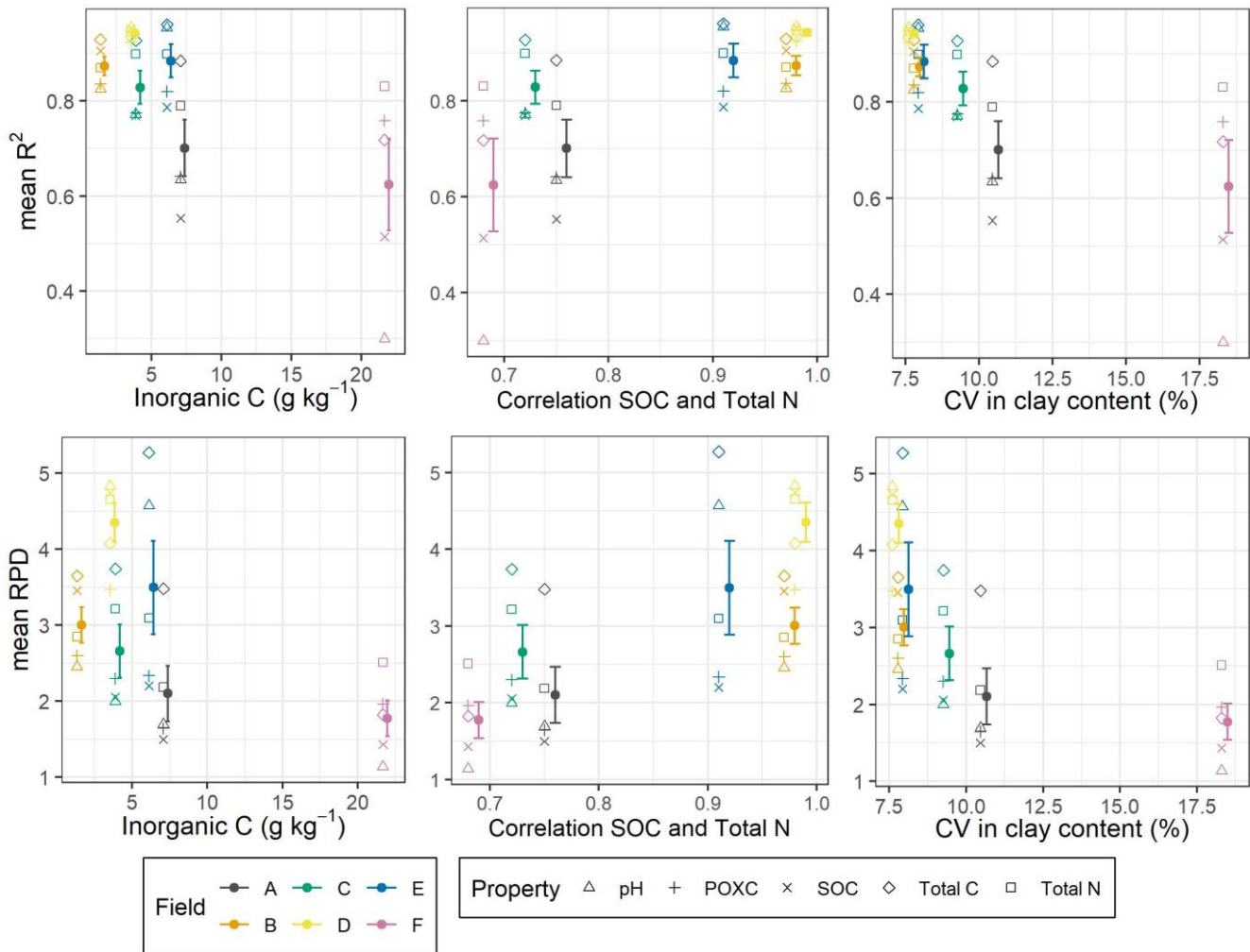



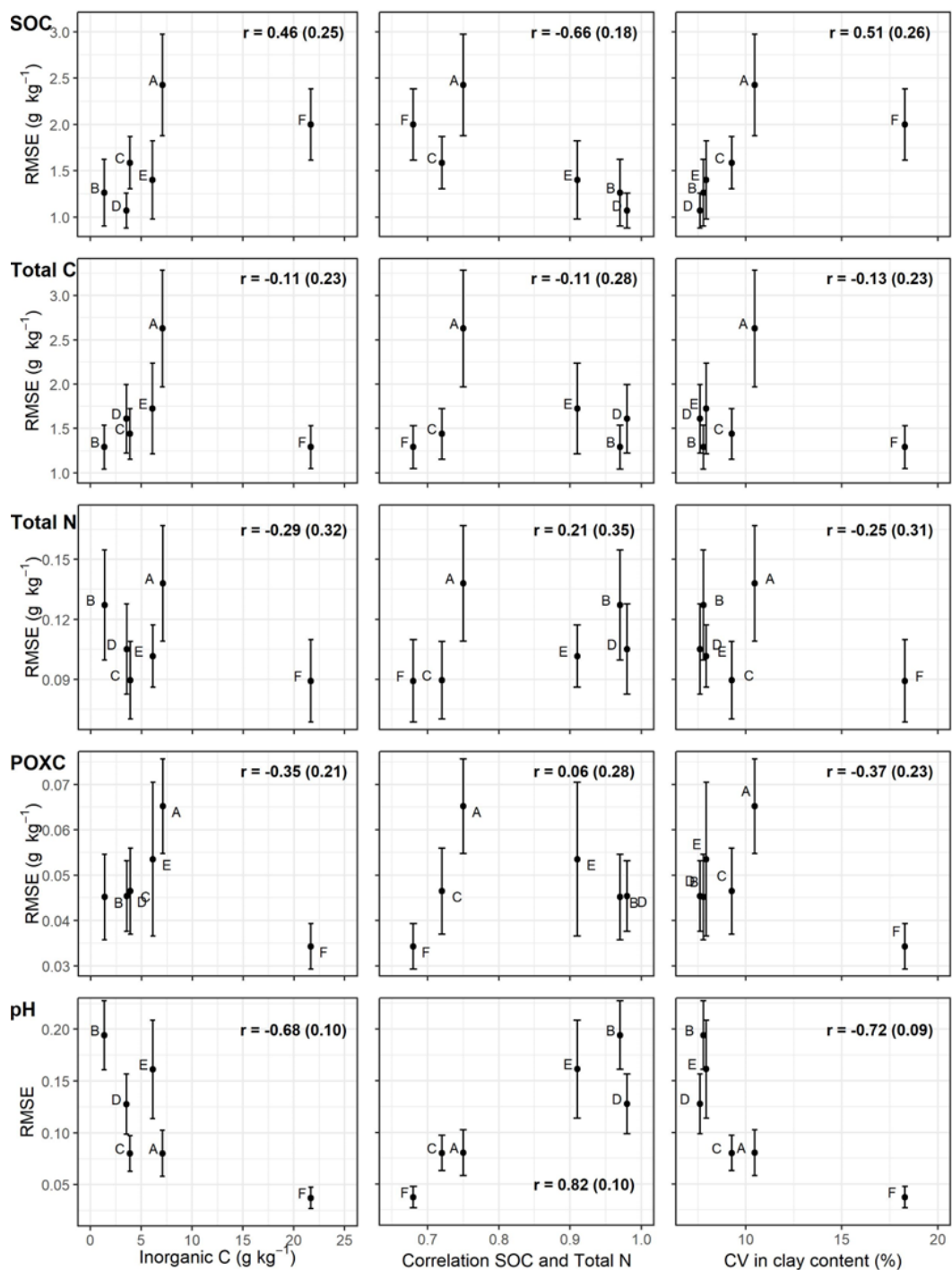
Figure 5: R^2 and ratio of performance to deviation (RPD) from the local models for SOC, total C, total N, POXC and pH aggregated (mean and standard error) per field (A-F) in dependence of mean inorganic C content, Pearson's correlation coefficient between SOC and total N and coefficient of variation (CV) in clay content. The error bars represent standard deviations across the repeats in the cross-validation.

635

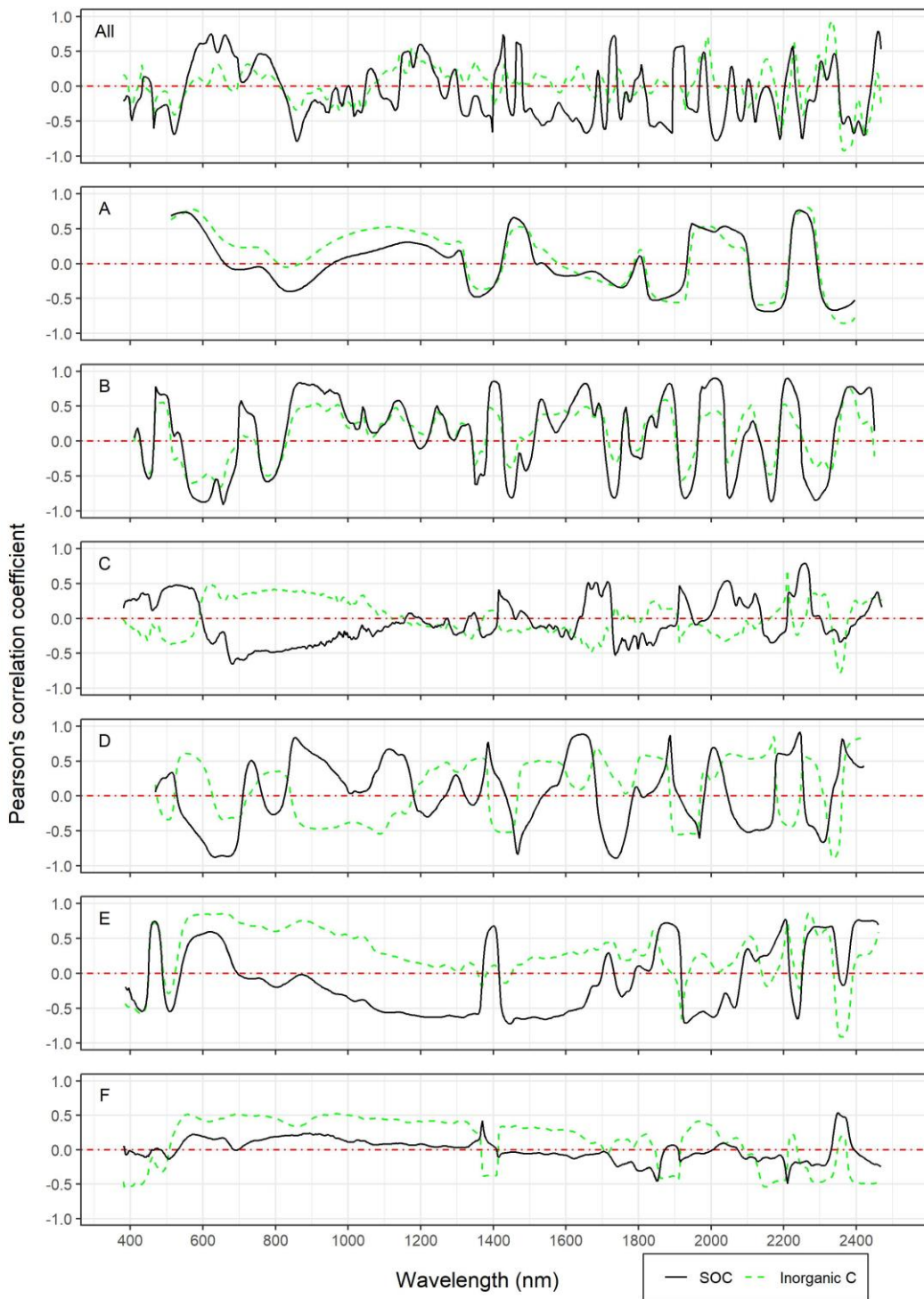
Figure 6: Root mean square error (RMSE) for the five target properties (SOC, total C, total N, POXC and pH) and each field (A-F) in dependence of mean inorganic C content, Pearson's correlation coefficient between SOC and total N and coefficient of variation (CV) in clay content. The error bars represent standard deviations across the repeats in the cross-validation. Pearson's correlation coefficients are indicated as mean and standard deviation (in parantheses) of a Monte Carlo simulation

640

Figure 7: Correlation graphs between spectral variables at each wavelength and SOC as well as inorganic C for the combined dataset (All) and the individual fields (A-F). The spectra were pre-processed according to the chosen models for SOC.

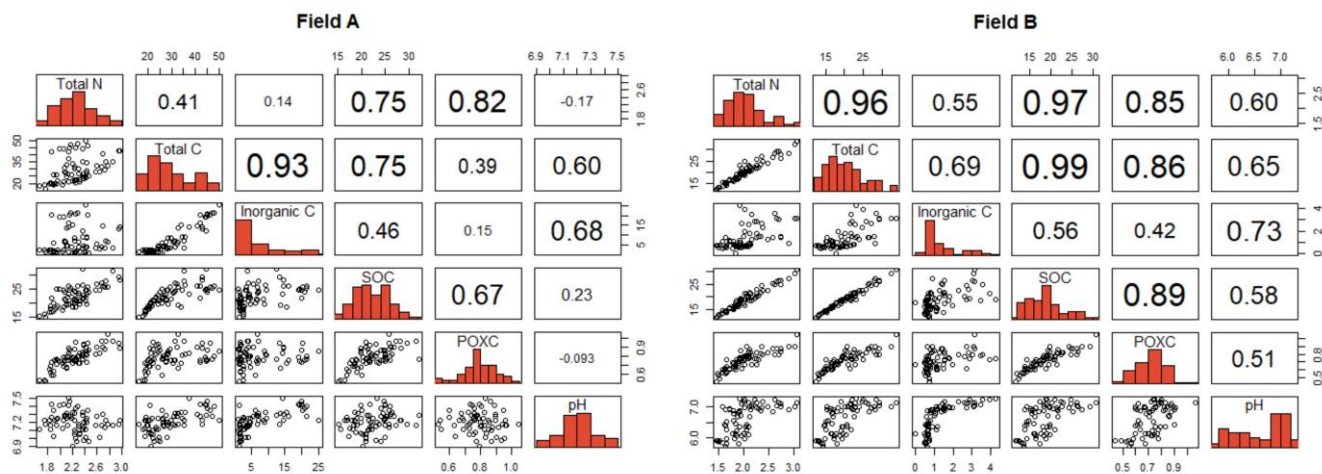



Figure 8: Pairwise correlation matrices between target soil properties and inorganic C for a field with weak (Field A) and field with strong correlations (Field B) between the target variables. The correlation matrices for all fields can be found in the supplementary material (Fig. S4).

Code and data availability:

Data and R-codes will be made available upon publication on a Zenodo repository.

Author contributions:

All co-authors conceptualized the paper. SO collected the data and conducted the analysis with the help of LS. SO wrote the original draft and all co-authors participated in the writing and editing of the manuscript.

Competing interests:

The contact author has declared that none of the authors has any competing interests.

Acknowledgements:

The authors gratefully acknowledge Daniela Fischer and Patrick Neuhaus for their engaged support during the lab work. We also warmly thank Philipp Baumann for sharing his knowledge about the handling and modeling of spectral data. Lastly, we would like to thank the two anonymous reviewers for their detailed and very constructive comments and suggestions.

Financial support:

This study was funded by the Land Systems and Sustainable Land Management Group from the Institute of Geography at the University Bern.