Dear reviewer

We thank you for your time and valuable comments. Please find our replies displayed in blue color below each comment.

Kind regards
Mareike Ließ

Anonymous Referee #3
General comments
The study by Ellinger et al. represents a under-researched topic with high practical relevance for soil spectroscopy applications. An exhaustive set of uncertainty factors contributing on the model error was examined, using a full factorial design. The chosen title is attractive and is well aligned the objective and the performed statistical analyses. Statistical model development and assessment was done using a state-of-the-art cross-validation technique. The assessment of uncertainty comprises a combination of two sampling strategies including their combination, four spectral preprocessing methods, and spectral and analytical data set combinations with different degree of averaging random noise. This discussion article is well structured and describes the technical aspects in precise and easy understandable plain language. Although covering rather fundamental spectroscopy modeling research, the spectroscopic calibration of samples from a long fertilization trial embeds the uncertainty analysis into a realistic and interesting application context (soil monitoring). This study backs up the current practice of measuring several sample or sub-sample replicate and subsequent spectral averaging, which can improve model performance in many cases.
Reply: Thank you.

Further, the data set and the error analysis is of particular interest for the soil science community because it quantifies the contribution of random analytical reference method errors on spectral predictions. This paper would comprise an even more valuable scientific contribution if the authors elaborated more on this particular uncertainty component.
Reply: We have rewritten large sections of the introduction and results section.

The inclusion of the sampling design as a factor together with the other varying uncertainty components requests more concentration from the reader to understand the experiment and read the results. In order to simplify the results and to make the message more concise, the authors could focus on the scenario with the combined “A” and “B” sampling strategies, and move the results of “A” and “B” alone to the appendix.
Reply: We understand the reviewer’s concern about the complexity of the paper. However, as the sampling design is an important aspect of this paper. We, therefore, refrain from shifting the corresponding results to the appendix.

The model tuning (number of PLSR components) results would be important for the interpretation of model performance under the uncertainty scenarios, but these are missing. These together with a more detailed discussion of error patterns not following the general trend or expectations would help to validate and explain the highly variable results. Further, light scattering effects as important source of spectral model error, and their dependence on soil composition and texture could be discussed, taking into account the applied spectral preprocessing methods. Addressing the specific comments, this paper will be a scientifically sound and valuable soil spectroscopy case study.
Reply: The PLSR components vary largely in dependence on the pre-processing method. The information on the number of selected components, therefore, did not result very informative. Furthermore, it

distracts the reader from the main message. We, therefore, refrain from including it in this publication. We have adapted the results section concerning the uncertainty scenarios. We have also extended the introduction section to elaborate on the various sources of uncertainty.

Specific comments

The suggested references are listed in the bottom of this review.

Abstract

Instead of giving a general description of model building, more specific details on the error propagation experiment would show the importance and the quality of the carefully conducted statistical experiment. The abstract lacks quantitative results about the contribution of tested uncertainty factors on model performance.
Reply: We have rewritten large sections of the introduction and results section to go more into detail on the error propagation experiment. Quantitative results with respect to the contribution of uncertainty factors on model performance were added to the abstract.

Introduction

The introduction reads fluently and is of adequate length. The importance of soil organic carbon for soil functioning and frequent measurements provides a good motivation to introduce the methodological relevance of the present study. The reader of this journal is likely already familiar with the role and importance of soil organic carbon an its assessment in soils; therefore, this part can be condensed. There is some statements that can be assumed common knowledge. As an example, the following sentence in l. 42–44 requires no references: "The precise monitoring of SOC on a LTFE with conventional lab analysis is labour-intensive and expensive (Adamchuk and Viscarra Rossel, 2010; Loum et al., 2016) as it requires the analysis of a rather high amount of samples." The authors clearly state the motivation of using soil spectroscopy. Shortening the paragraph on soil organic carbon, there is space to briefly explain principles that enable sensing of carbon by infrared spectroscopy (e.g. relationship between functional groups in soil constituents, absorption of electromagnetic radiation and vibration of bonds in the infrared range, and soil properties, and how these relationships were used for statistical modeling).
Reply: The section on spectral sensing was moved from the methods section to the introduction. However, we refrain from shortening the section on SOC as it is important to understand the context of the Vis-NIR application: Soil monitoring.

The theoretical foundation of uncertainty in modeling is rather sparse and short in relation to the general introduction into concepts of soil organic carbon and advantages of spectroscopy (see also comment of anonymous referee No. 2). The authors are advised to add more key concepts and terminology of uncertainty in the context of predictive modeling generally, and soil spectroscopy modeling applications specifically. It is worth mentioning that measurement and prediction errors can be separated into systematic (bias) and random errors (uncertainty, precision). There are several sources of uncertainty in the model predictions, such as predictor (spectra) measurement errors, response (chemical laboratory measurements) errors and model errors (related to model instability in model parameter estimates or model structure). For analytical data, random errors are most relevant. Further, both bias and variance contribute to the uncertainty in generalization error estimate (here RMSE), which is another source of uncertainty.

Reply: We have extended the section on the various sources of uncertainty in the introduction. However, as we do not distinguish between systematic and random errors in this study, we refrain from referring to them in this explicit way.

The authors mention and stress importance of resampling strategy, which was comprehensibly explained. To complement, a link to the conditional model error could be made (for example citing Beleites et al. (2005)).
Reply: We have included a reference to Beleites et al. and Molinaro et al. in the context of resampling.

Some more details that are more relevant for the authors' experiment can also be added to the existing section in Material and Methods. For diffuse reflection infrared spectroscopy, scattering effects are particularly important spectral noise factors, that are relevant under the chosen experimental setup and the objective. Therefore, they merit particular mention.
Reply: We have added two references concerning sensor noise and scattering in the introduction section (Schwartz et al., Pilorget et al)

For a general description of model uncertainty analysis, e.g. Jansen and Michiel (1998) provide a good reference.
Reply: Thank you. We added the reference.

l. 55: Wetterlind et al. (2013) recommend general strategies for spectral measurement and modeling; Spectral averaging is also recommend. Therefore, this would be an important message and reference to add to the existing ones. Further, Wetterlind et al. (2013) highlight the effect of the sampled area and advise to perform replicate spectral sampling for small areas.
Reply: We refer to spectral averaging and repeated scans in the introduction and discussion section. We refer to the standard measurement protocol suggested by Pimstein et al. and, therefore, refrain from citing Wetterlind et al. in this context.

Material and Methods

The model of the CN analyzer is not described.
Reply: added

l. 104f: "C measurements were taken as organic carbon due to negligibly small carbonate contents." The authors are asked to provide quantitative statement (lower than xxx % C.).
Reply: Carbonate contents were below detection limit.

There is no mention of whether or how many scanning replicate measurements were internally averaged (spectrometer setting) prior taking spectrometer readings of the sub-sample and rotation replicate spectra (further noise averaging).
What is the name or composition of the material was used as a white reference (name manufacturer if composition not known)?
Reply: The missing information was included

The description of the preprocessing techniques is a bit too long and detailed. To keep a focus on the main topic, a brief description in one or two sentences, a citation to the original publication and maybe some soil spectroscopy case study where the respective techniques was applied, suffice. Spectral preprocessing

techniques have been well described and researched in chemometrics, and applied in various other scientific discipines and industry over the last decades.
Reply: As we also assess the impact of the pre-processing method on the error measure, we think a detailed description is necessary for understandability.

The chosen resampling strategy is particulary well suited when data is scarce and variability is high, but can be recommended in general. This study can serve as an exemplary resampling setup for the soil spectroscopy community (repeated nested group k-fold cross-validation for model parameter tuning and evaluation). The approach is also particulary consisely described. Repeated 10-fold cross-valiation is a practical and widely used cross-validation strategy to reduce uncertainty in performance estimators.Therefore, adding another reference to earlier applied predictive modeling literature that study effects of resampling strategy on the estimation of model evaluation metrics is advised. For example, Molinaro et al. (2008) is a suitable reference. The description and the illustration in Fig. 6 refer to a nested or double cross-validation, where the inner resampling layer comprises parameter tuning and the outer layer is used for model evaluation, which is used to avoid selection bias in parameter tuning. The authors should mention the nested or double cross validation terminology and also reference key literature. Varma et al. (2006) is suggested as a reference here.
Reply: We specified the applied cross validation as nested approach and added Varma and Simon as well as Molinaro et al.

“Although the samples were sieved and homogenised before taking measurements, within-sample variability still had an influence on the model outcome. In general, the variability of the samples depends on the soil treatment and possibly also on the origin of the samples (e.g. agricultural or forest soils).” The influence of scattering effects and major importance of texture should be mentioned and backed up with literature in this place.
Reply: Reference to soil treatment and scattering effects was made in the introduction. We refrain from referring to soil texture as we are at within-field scale and do not have a pronounced textural variability in our dataset. A reference to sample origin is included in the discussion section.

Results and Discussion

General considerations on the reviewer’s comments: Some of the below comments may cast a rather sceptical view on the authors’ explanation of the results. However, the experimental conditions (in a statistical sense) are not necessary such that there is major weak points of the results. The authors are kindly asked to provide a response or some more results or explanations as outlined below (appendix). The intention is to challenge the author’s hypotheses. There are unfortunately no means to compare these results to other similar studies in the soil spectroscopy literature. Nevertheless, the spectroscopy community has been following the general recommendation to average noise from replicate measurements to obtain good performing and robust models. This study showcases some "surprising effects". There are no consistent patterns that clearly show beneficial effects of averaging and removing spectra under all conditions. Some inconsistent patterns many arise from interaction in contributing factors.
Reply: We have adapted the results and discussion section and also refer to these unexpected effects.

A repeated nested 10-fold cross-validation guarantees that parameter selection is unbiased, but is prone to overfitting if sample grouping is only respected in assessment and not in tuning layer, in combination of multiple spectral and/or analytical replicate data set that are used for modeling.

Reply: We have added a corresponding sentence in the conclusions: “We are aware that the consideration of stratified group CV only for model evaluation but not for tuning might impair model performance as suboptimal model parameters might be selected. This will be adapted in future studies.”

How many PLSR components were used in the final models? What was/were the frequency/frequencies of respective best number of components across the folds and repeats?
Reply: The PLSR components vary largely in dependence on the pre-processing method. The information on the number of selected components, therefore, did not result very informative. Furthermore, it distracts the reader from the main message. We, therefore, refrain from including it in this publication.

For Savitzky-Golay (method 1) and the Norris gap derivative (method 4), the RMSE in model set 101 is considerably higher than in model set 100 as well as all other sets, whereby 3 times 6 identical spectra instead of 3 times an identical average spectrum are used (Fig. 8). Could it be that this is the result of a deleterious preprocessing effect due to missing sensor jump correction (see Fig. 3)? Noise enhancement may occur under certain preprocessing strategies such as the gap derivative (see also comments in next paragraph for further interaction possibilities).
Reply: We have corrected for the sensor jump and rerun the models. Figures 8 and 9 were adapted accordingly. Still models build on $dataset_{101}$ in some cases perform worse than models build on $datsets_{100}$. We refer to this peculiarity but have no explanation. As nested repeated 5-fold CV resulted in similar model results, we have replaced 10-fold CV by 5-fold to save computation time, the same applies for the number of components tested, which was reduced to 30.

Fig. 5 shows that the spectral variation largely manifests as offset variation. Savitzky-Golay smoothing with no derivative does not remove offset errors. This is worth a discussion. Assuming identical resampling sets among preprocessing methods within a data set, there seems to be a systematic error component in the spectra in data set A for these cases. The authors are encouraged to recompute results after correcting for sensor offsets, or they should at least consider possible explanations for clearly poorer model performance in the discussion.
Reply: We have corrected for the sensor jump and rerun the models.

How do the authors explain such a drastic error increase in the data situation from 3 times 6 different averaged spectra to 3 times one final average spectrum per sample when using 3 analytical replicates? Why does this occur only when adding replicate analytical measurements to the modeling process, but not when averaged analytical measurements are used?
Reply: We have added this aspect in the results section: “It is not surprising that the dataset of 3 SOC replicate measurements with 1 averaged spectrum ($Dataset_{100}$) results in lowest model performance, as the within-sample variance concerning SOC cannot be explained by the contained predictor information, the input data uncertainty propagates through the model building process.”

The nested cross-validation can yield different optimal number of component for each fold, which can be justified such that different data situations comprise different optimal model parameters with respect to performance. The nested or double cross-validation scheme in Fig. 6. shows the model validation procedure in the outer cross-validation loop within the left partition set and the model tuning resampling procedure within the right partition set. The authors mention that k-fold cross-validation was done by assigning the entire set of replicate spectra of respective samples in either fitting or hold-out sets. The authors don’t explicitly state whether this grouping by sample is done for the both model evaluation k-fold cross-validation layer and model tuning cross-validation layer, or just in either one of them. This lets room for ambiguous interpretations of the experiment and confounding factors. The R function

‘tuneControl‘ used for the tuning resampling, but the ‘groupKFold()‘ function from the caret RÂ˘apackage – which splits data based on a grouping factor – is not mentioned. Namely, focusing on the results depicted in panel "c1"Â˘aof Fig. 8, the prediction error decreases when less sub-sample replicate spectra are included in the modeling. Assuming that simple 10-fold cross-validation without grouping, the tuning layer would suffer from a data leakage from fitting folds into the respective assessment folds when replicate spectra are present. Data leakage provides biased tuning results and can select very high number of components. The maximum number of PLSRÂ˘acomponents was set 40, which represents severe over-fit with 100 rows and so many predictors (“large p small n"" problem). The reader cannot interpret if there was variability of the number of finally selected components for the different resampling sets. The authors should therefore provide a ncomp final tuning results table in the appendix if they were variable among the sampling sampled data set, model set and preprocessing method combinations. In worst case scenario where the outer assessment resampling is grouped, but the inner tuning resampling lacked sample grouping, presence of more replicate spectral measurements per sample can yield too high ncomp and too adaptive models, This would give an alternative explanation for poor performance in the outer assessment when all replicate spectra are used for modelling, in comparison to model set011 and set001 with increasing degree of spectral averaging and less likely too adaptive PLSRÂ˘amodels.
Reply: Only the outer CV cycle of the nested repeated k-fold CV approach considers stratified group CV (but: spectral replicate measurements and scans per sample were always assigned to the same fold!). This was specified in the corresponding methods section. Of course, considering stratified group CV also for model tuning might improve model performance. As we used the inbuilt function of the Caret package, this was not possible. We are not aware of any publication that actually implemented stratified group CV in parameter tuning. We will, however, implement this in future studies as mentioned in the conclusion section. The PLSR components vary largely in dependence on the pre-processing method. The information on the number of selected components, therefore, did not result very informative. Furthermore, it distracts the reader from the main message. We, therefore, refrain from including it in this publication.

Using the full set of analytical carbon measurement replicates can have generally lower performance. Directly addressing these aspects is an outstanding achivement of this study. The authors are highly encouraged to extend the discussion about effects of analytical uncertainty, taking into account the considerations of anonymous referee No. 2..
Reply: We have further elaborated on this aspect.

To sum up, the authors are kindly asked elaborate and comment on interactive effects between the sources of errors and the resampling, thereby also critically address confounding effects as hypothesized above in the author’s response to the review. Did the authors implement the grouped k-fold procecedure for both inner an outer cross-validation layers?
The authors could move Fig. 3 to the results and further illustrate differences in preprocessed spectra for replicate spectral measurements (e.g. one example spectrum).
Further, texture might explain different model performances for data sets “A”Â˘aand “B”. Sandy soils are known to confound increases in reflectance typically also found for increases in soil organic C (see e.g. Stevens et al., 2013).
Reply: These aspects were addressed in our replies to the specific comments.

The use of different preprocessing techniques has already been exhaustively discussed in the soil spectroscopy literature. The performance of different preprocessing methods depends on the data context, as stated in the last sentence of paragraph l. 347–349. Thus, such a comparison to other studies does not make sense and brings no added value to the reader. The authors are advised to remove the

comparison. To mention that the effects of signal processing methods and associated parameters on model performance are study and data specific is sufficient (incl. references).
Reply: We are not aware of any study that actually quantified the effect of spectral pre-processing on model performance and, therefore, refrain from deleting it from our study.

Conclusion

“Autocorrelation between calibration and validation sets”. It is valuable for the soil spectroscopy community that the present study considers and stresses data grouping effects in resampling, here repeated measures, and accounts for those in the crossvalidation procedure. Ignoring such grouping factors can result in over-optimistic estimation of the model performance due to leakage of predictive relationships from the modeling into assessment sets. The scientific community would benefit if authors could name the strategy using the terminology "group k-fold cross-validation".
Reply: The aspect was emphasized in the conclusion section.

Technical corrections

l. 86–89: “Categorical and continuous data first entered a factor analysis with mixed data (FAMD) performed with R package FactoMineR (Lê et al., 2008) to allow for further joint analysis. For design ’A’ the LTFE plots were then grouped by a k-means cluster analysis. R package NbClust (Charrad et al., 2014) automatically determines the optimal number of clusters making use of 30 indices.” How many factors from previous factor analysis with mixed data are retained and used for k-means clustering? How much variance do these factors and continous variables explain.
Reply: The FAMD was applied to decorrelate the data. All factors were retained.

l. 127–132: The outlier analysis is in section 2.4, but is not considered as preprocessing.
This part needs either a separate section or a generic data analysis section.
Reply: We refrain from including an additional section for only two sentences. As the outlier removal is the first step before applying any spectral pre-processing, we do not see why it should not stay in this section.

l. 263f: “The violin plots of all three soil sample sets do not resembles the archive violin plot very much.” typo: "resemble"
Reply: corrected

l. 344: “The simple first derivative (d1) performs poorely because it increases noise as stated, but there is no tendency to overfit related to this particular preprocessing technique (rather consequence of improper model resampling and tuning setup in combination with adaptive models)”; typo: “poorly” .
Reply: corrected

l. 365f: “The impact of the different pre-processing methods showed clearly in this study, but it may be different when using other data sets, though.” This sentence needs to be more precise; the authors are kindly advised to use "impact the different preprocessing methods on model performance..." or mention variable performance results or similar.
Reply: The section was rewritten.

l. 239f: “In this study, the partition of the spectral data into 10 folds for 10-fold CV had to be done very carefully, as in some cases multiple spectra existed for one sample.‘

Expression ”very carefully“ needs to be rephrased in scientific language, as well as “some cases” (2/3 of the cases).
Reply: adapted

Figures

General comment for panel figures: Experimental labels are hidden within the plot area and are better placed on top left of each subplot to facilitate the reader’s visual perception.
Reply: adapted

Fig. 3: The spectra have not been joined correctly in the sensor jump region (900nm).
On the left side of the sensor shift there is a small peak which appears to be identical on the right side.
Reply: Thank you. We corrected for the sensor jump.

Fig. 5: Zooming into the range with highest replicate spectral variation would help to discriminate single replicate spectra and help the reader to visually assess the type of errors in the spectra (random noise vs. systematic offset).
Reply: Thank you. The figure was adapted, accordingly.

Fig. 9: The figure quality needs to be improved for print. Plots should be in a vector format. Increase the font sizes for axis labels. Also, mathematical expressions should be separated by full or half spacing. Use minus sign or alternatively en dash instead of hyphen for minus. R-squared and RMSEÂˇacould be placed below each other for better readability. Hollow circles or transparency, and bigger point symbols to deal with overplotting are recommended. Letters "a/b" in panel labels should be capitalized because the sampling designs are abbreviated as "A/B" elsewhere.
Reply: adapted

References

Beleites, C., Baumgartner, R., Bowman, C., Somorjai, R., Steiner, G., Salzer, R., & Sowa, M. G. (2005). Variance reduction in estimating classification error using sparse datasets. Chemometrics and Intelligent Laboratory Systems, 79(1–2), 91–100. https://doi.org/10.1016/j.chemolab.2005.04.008

Jansen, M. J. W. (1998). Prediction error through modelling concepts and uncertainty from basic data. Nutrient Cycling in Agroecosystems, 50(1), 247–253. https://doi.org/10.1023/A:1009748529970

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301–3307. https://doi.org/10.1093/bioinformatics/bti499

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

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

Wetterlind, J., Stenberg, B., & Rossel, R. A. V. (2013). Soil Analysis Using Visible and Near Infrared Spectroscopy. In F. J. M. Maathuis (Ed.), Plant Mineral Nutrients (pp. 95–107). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-152-3_6