Reply to Reviewer 1

12.09.2023

1 General answer

We thank you for your detailed feedback and very constructive comments. We appreciate that you see value in our study and approved the handling and modelling of the spectral data.

You raised two major concerns and several specific comments that we address in our answers to your review. For more clarity between our answers and your comments, we have inserted your original statement in blue color.

We hope we addressed all comments to your satisfaction and thank you for your suggestions that help us to further develop the manuscript.

On behalf of all co-authors,

Simon Oberholzer

2 Concern 1: The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed

We thank you for these detailed explanations which show the problematic of this analysis. We understand that our analysis about dataset variability and performance of spectral models has limitations because of the interrelation between coefficient of variation (CV) and the model performance parameters (R^2, RPD, PRMSE). We will remove research question 3 from the manuscript (chapter 3.5, 4.4 and 4.5) as you recommended. Moreover, we will put more focus on research question 4 about the controls of predictive performance of spectral models. We thank you for summarizing that topic in four bullet points:

1. Focus on estimates of absolute predictive performance.

2. Consider target variable identity.

3. Understand the factors which cause correlations between spectral variables and a target variable and the factors which confound or mask these relations.

4. Understand what errors are caused by extrapolation.

We will shift the focus from R^2 and RPD to RMSE, which also means that absolute model performance must be discussed in more detail for each property separately. As you suggested, we will expand the analysis to identify soil properties that influence the predictive model performance (section 3.5). So far, we showed the influence of mean carbonate content on model performance. We will expand this analysis and assess the impact of a) the variability of carbonate content and texture, and b) the correlations between soil characteristics (carbonate and texture) with the target variables on the performance (absolute prediction error) of the spectral models. This will provide insights into which factors mask the correlation between spectral variables and target variables.

3 Concern 2: Data and code must be available to interpret, replicate and build upon the findings reported in the article and should be published.

We will add a supplementary file containing the R codes for the spectral modelling where all steps can be followed.

4 Specific comments

Comment 1:

13 to 15: "… general models (combining all fields) for organic carbon, total carbon, total nitrogen, permanganate oxidizable carbon and pH using partial least squares regression. 24 out of 30 local models showed an accurate or even excellent performance (ratio of performance to deviation (RPD) > 2) …". I think the statement does not provide any useful information because it is not used to compare performance of models, but to make a standalone statement for one specific model (compare with point 4 in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" of my review).

Since we have 6 local field models for each of the 5 properties and want to give a qualitative judgment about their relative performance, we use the RPD and the classification system proposed by Chang et al. (2001) and Zhang et al. (2018). Thereby we mainly want to give an overview about how many of the developed models showed a good performance. However, we agree with the you that the RPD should only be used when comparing multiple models of one property and will use it in such a context only. The RMSE as an absolute measure is the most important parameter and we will discuss it in more detail for the individual models, which can still be accurate even though they were classified as not satisfying based on RPD.

Comment 2:

15 to 16: "… and the root mean square errors (RMSE) of prediction were, except for pH, maximum five times higher than the lab measurement error." Whilst this statement is true (and RMSE is useful as a measure of model performance for standalone statements on model performance), I am not sure whether framing the comparison of model and lab RMSE only in terms of a ratio is useful (this also is the case in the results and discussion sections, for example) because one reason why the RMSE for pH is five times higher than the lab measurement error is that the pH lab measurement error is so small (see Fig. 3). Of course, all this makes only sense when one knows, what absolute error is tolerable.

In this statement we mainly want to show how much accuracy is lost when using local spectral models instead of conventional lab measurements. However, we agree that more detail about the lab measurement accuracy for each soil property is needed to make a fair comparison. Therefore, we will discuss the comparison between RMSE of the local models and the lab measurement error for each soil property separately. We will also discuss for each property what would be a tolerable prediction error.

Comment 3:

68 to 70: "However, it remains unclear what an optimal variability of a soil property in a project of local extent would be to achieve a high measurement accuracy in absolute values (low RMSE) but also

relative to the range of the data (RPD).” Just to link this to my comments in section “The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed”: I think this question is misguided because it does not depend on the variability of the target variable how good a predictive model is (except that the variability of the target variable to predict should be larger than measurement errors), but on how well spectral variables are related to the target variable and masked by others. Of course, often a larger variability of the target variable implies more masking of spectral variables related to the target variable by other variables, but this larger variability of the target variable is not the cause.

We will remove question 3 (see answer to Comment 1).

Comment 4:

84: I think it makes sense to replace “measurement accuracy” by “measurement error”.

We thank you for the suggestion and will replace ”accuracy” by “error”.

Comment 5:

124 to 126: “To estimate the measurement error of SOC we took the sum of the standard deviation of the total C measurement with the standard error of the inorganic C measurement because inorganic C measurement were for all samples done in triplicates.” Shouldn’t this rather be $\sqrt{\sigma_1^2 + \sigma_2^2}$ (instead of $\sigma_1 + \sigma_2$), where σ_1 is the standard deviation of the total C measurement and σ_2 the standard error of the inorganic C measurement?”

We thank you for pointing this out. We will correct the calculation of the lab measurement error of SOC.

Comment 6:

145 to 147: “The pre-processing techniques … led to around 100 meaningful combinations that were tested in model building and the final pre-processing option was selected based on the lowest RMSE.” Have you selected the model with the lowest computed average RMSE, or have you applied the one-sigma-rule also when comparing RMSE between different preprocessing workflows? If it’s the former, redoing this analysis while applying the one-sigma-rule would make sense to avoid overfitting and to consider the uncertainty in the RMSE estimate when discussing which preprocessing methods performed well. In particular, I expect that within one standard deviation multiple preprocessing workflows performed equally well and one should decide that not enough information is available to pick prefer a specific one of those.

Yes, we did apply the one-sigma-rule when choosing the optimal preprocessing and you are right that many of the preprocessing methods worked similarly well. We will provide more details in the manuscript and add examples of model performance with different preprocessing workflows in the supplementary file.

Comment 7:

149 to 151: “Since all soil properties showed a limited skewness (see Table S1 in the supplementary material) that was always in the range of $-2 \leq skew \leq 2$ which was proposed as acceptable to assume a normal univariate distribution (George and Mallery, 2010) we consider the application of PLSR appropriate, especially since it is robust to minor deviations from a normal distribution (Goodhue et al., 2012).” Some of the statements are not appropriate: (1) The skewness limits are a bad way to test whether a distribution is a normal distribution (See R code 1 below where I simulate data from a

distribution which obviously is not normal and nevertheless has a skewness within the specified boundaries). (2) The data itself do not have to follow a normal distribution for PLSR to provide valid results (See R code 2 below where I simulate data from the same non-normal distribution and show that a PLSR model with good performance can be computed on these data).

We thank you for raising this important point and providing the example R codes. We will adapt this section accordingly and write that PLSR is more robust with normal distribution, but that normality is not a mandatory requirement.

Comment 8:

166: "… we calculated the mean Euclidean distances between all samples …". I assume the Euclidean distance was computed using the preprocessed spectral variables?"

We calculated the Euclidean distances using the raw spectra since in the beginning it was not clear which pre-processing method would be suitable. However, we agree that it makes sense to use the preprocessed spectra. We will adapt the distance calculations and use the preprocessed spectra, which we used most frequently for the trained models (Resampling interval = 3, Savitzky-Golay filter and multiplicative scatter correction)

Comment 9:

174: "Since we used a cross-validation approach on the field scale, all models showed a very small bias." Please provide evidence for this claim, e.g., in Tab. 2 or in the supporting information.

We will include the bias in Table 2 to support this statement.

Comment 10 – 13:

We agree on all these minor comments and will implement them as you suggested.

Comment 14:

221 to 222: "… we analyzed the influence of mean carbonate concentrations, soil texture and field size on the model metrics but only the mean carbonate concentration showed effects and is therefore presented in the results." This description needs to be much more detailed: (1) How did you analyze this (regression, correlation, in case of regression, which distribution was assumed for the target variable? (2) Which models were computed (what variables were included)? (3) How (based on what criteria) were variables included/excluded? (4) Was this analysis performed separately for different target variables (as suggested in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" above)? (5) Is the uncertainty in the estimated RMSE small enough to not affect your results and have you checked this? In addition, please provide results for all models which were computed to support your claim that "only the mean carbonate concentration showed effects" (this can be shown in the supporting information). If this is an option for you, I strongly encourage you to expand this analysis, perhaps combining this with the analysis of the variable importance, because this is in my opinion the right direction to better understand the controls of model predictive performance (compare with my comments in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed").

We thank you for raising these important questions. With additional analyses we will investigate the effect of additional variables on the absolute model performance / RMSE. We will also provide more details on how variables were selected and how their correlation with the modeling errors was

assessed. Furthermore, we will add more information on variables, which did not show a significant impact on model errors and where therefore excluded from the manuscript. The stronger focus on the VIP values in the analyses will furthermore give insights on some potential interferences of soil properties in the vis–NIR range. As shown in Figure 4, for the models for each soil property, the important wavelengths are very site-specific, which can explain their varying model performances. We plan to check for correlations between soil properties (i.e., carbonate, texture) and spectral variables (as done for example by Conforti et al. (2018)) and compare them with the VIP analysis. This analysis might give some insights which soil properties might mask the spectral features of the target variables. Please find more details regarding this comment above (Concern 1) and answers to the five specific questions below:

1. Since we have six fields, we were always looking for linear relationships between site characteristics and parameters of model performance because it is very difficult to assume a different distribution with this number of locations. Thereby we mainly focused on linear correlations. We will add these important but missing details to the manuscript.
2. We explored linear relationships between field size, soil texture and carbonate content and model metrics (RMSE, R^2 and RPD). The manuscript will be completed with the information on these analyses.
3. We selected the available variables which we assumed had an impact on model performance.
4. We performed this analysis for all variables combined as well as for each target variable individually which we will specify in the manuscript.
5. We did not consider the uncertainty in RMSE for this analysis, but we will do include it in our revised version (as already suggested above; see Concern 1).

Comment 15:

225 to 226: Please specify the versions of the R packages as this can be important to reproduce your analyses.

We will indicate the R-package versions.

Comment 16:

246: “There was no pre-processing combination that proved to be suitable for all models of the same field (Fig. S2 in the supplementary material) or of the same soil property (data not shown).” Please provide evidence for this claim, e.g., in the supporting information.

We will give examples of model performance with different pre-processing methods in the supplementary material.

Comment 17:

268 to 269 and elsewhere: “The field-specifically calculated R2”. Maybe “field-specific R2” is more concise?

We thank you for this suggestion. We will follow this advice.

Comment 18:

283 to 284: “bit higher for POXC and substantially higher for pH (Fig. 3).” Do you mean a bit higher for pH and substantially higher for POXC? The figure does not seem to fit to the text, at least if this difference is meant in terms of absolute values (~0.15 to 0.25 pH units versus >20 POXC units) (compare also to point 2 in this section of my review).

That was a writing mistake. We will correct it accordingly.

Comment 19:

Section 3.5: Are uncertainties in the PRMSE and RPD negligible for the results of your analyses? (Please note my comments in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" above. I just wanted to mention that this point is not clear).

We will indicate the uncertainties of RMSE and RPD and take them into account when we report and discuss model performances.

Comment 20:

All analyses of the relation of R2, PRMSE, RPD versus some other variable: You've described that you used a Weibull distribution for R2 since R2 must be in [0, 1]. However, prediction intervals shown in Fig. 5 and 7 contain values larger than 1 and thus do not comply with this assumption. Similarly, PRMSE and RPD cannot have negative values, but prediction intervals for PRMSE in Fig. 5 and 7 contain negative values.

This comment concerns the analysis and figures about dataset variability and model performance that we will remove according to your recommendation.

Comment 21:

410 to 412: "Even though the local models of fields A and F had the lowest performance among all local models, they still showed, with exception of pH, approximate results which, depending on the research question might still provide useful information." I assume that "approximate" refers to the classification of RPD? In that case, I think the statement does not provide any useful information because it is not used to compare performance of models, but to make a standalone statement for one specific model (compare with point 4 in section "The analysis of how dataset variability controls model performance (research question 3) has major limitations and should be removed" of my review).

We thank you for the comment. We will, throughout the manuscript, use the RMSE for standalone statements and RPD for comparison between models (see above / Comment 1).

Comment 22:

423 to 424: "… we did not find a better model performance with increasing mean clay content …" Please provide evidence for this claim, e.g., in the supporting information.

We only looked at the linear correlation of the mean clay content with the model performance metrics, but we will expand this topic and provide more information on the methods as well (see above / Concern 1 and Comment 14).

Comment 23:

556: The URL "https://doi.org/10.2136/sssaj2002.6400" is wrong. It should be "https://doi.org/10.2136/sssaj2002.6400a".

We thank you for spotting this typo and will correct this DOI-link accordingly.

Comment 24:

698 to 699: "Lastly, we would like to thank the two anonymous reviewers for their constructive comments and suggestions." Thank you ;)

A sentence that was written to early but has proven to be very true.

Comment 25:

Fig. 1: Why are no density plots for carbonate concentration and soil texture variables shown here (or in the supporting information)?"

We agree with you, it would be helpful to have the density plots also for the carbonate contents and soil texture. We will include carbonate contents that in the figure as well. However, we only have texture measurements for composite samples per field analyzed with the standard sedimentation method. Additionally, we have soil texture data measured with laser diffraction method (LDM, Mastersizer 2000) in higher resolution (20 samples per field). The two methods are very comparable for sand content but the clay content was substantially lower and silt content higher with the LDM (also described in Ryzak and Bieganowski, 2011). We will scale the mean of the LDM measurements to the average of the sedimentation method while keeping the same relative standard deviation to obtain the heterogeneity of soil texture in the field. The higher resolution of soil texture will also be used to identify potential correlation with the target variables which might further show interferences with spectral features (see above; Concern 1, Comment 14).

Comment 26:

"Fig. 3: R2, RMSE, RPD, and the lab measurement error are estimates and have uncertainties which should be added as error bars to the figures."

We will add error bars for R^2, RPD and RMSE to the figures.

Comment 27:

"All figures (also in the supporting information) would benefit from (1) a higher image resolution, (2) larger legend keys, (3) thicker symbol lines or different symbol shapes such that it is possible to recognize better which points have which color."

We will enhance the resolution and the graphical presentation of the figures.

5 References

Chang, C. W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R., 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Science Society of America Journal 65, 480-490.

Conforti, M., Matteucci, G., and Buttafuoco, G., 2018. Using laboratory Vis-NIR spectroscopy for monitoring some forest soil properties. Journal of Soils and Sediments 18, 1009-1019.

Ryzak, M., and Bieganowski, A., 2011. Methodological aspects of determining soil particle-size distribution using the laser diffraction method. Journal of Plant Nutrition and Soil Science 174, 624-633.

Zhang, L., Yang, X. M., Drury, C., Chantigny, M., Gregorich, E., Miller, J., Bittman, S., Reynolds, W. D., and Yang, J. Y., 2018. Infrared spectroscopy estimation methods for water-dissolved carbon and amino sugars in diverse Canadian agricultural soils. Canadian Journal of Soil Science 98, 484-499.