Response to the reviewers: "Developing the Swiss soil spectral library for local estimation and monitoring" by Baumann et al.

https://soil.copernicus.org/preprints/soil-2020-105/

Letter of Response

Dear Prof. van Wesemael,

Thank you for guiding the review process of our research presented in this pre-print article. We thank the two reviewers, who carefully evaluated our manuscript and invested a significant amount of time to outline the aspects that we can improve. In our general comments and detailed comments (Authors comments; AC), we refer to the comments made by both of the reviewers (RC), and give more detail on how we would like to improve the manuscript.

Best regards,
Philipp Baumann, on behalf of all co-authors

1 Synthesis

AC: Reviewer one provided detailed comments marked in the manuscript (see below), but did not provide any synthesis.

RC2: The paper presents a Swiss MIRS soil library (>4000 samples collected at 0-20 cm depth; 17 properties considered), which is used for developing national prediction models with Cubist (decision and regression trees) and site-specific models with RS-LOCAL (which selects tailored calibration subsets from the library). Presenting such national spectral library is particularly interesting a priori.
AC: Thank you for this acknowledgement.

2 General comments

AC: Reviewer one did not provide any general comments.

RC2: However, methods for using such libraries represent important matter, but their description is sometimes too succinct (e.g. 10 words for describing how Cubist works, though it is complex and not very popular yet... while 10 lines are dedicated to the Mahalanobis distance, widely used in spectral analysis); and when not, it is sometimes difficult to follow, and clarification would be welcome. In contrast, rather evident considerations on the interest of such library are developed extensively.
AC: We thank for this valuable perspective, pointing out that we must take into account the readers' wide background extensively. There is inevitably a growing interest in developing and expanding spectroscopic libraries. Although the motivation and benefits behind the construction and the expansion of diverse soil spectral libraries seems obvious in light of soil quantitative data needed, we observe that the specific approaches to leverage them are still diverse; while fundamental working principles are very established and have long been tested. It appears that much of the currently innovative methodological research is centered around the following question: How can we achieve "more local" (for example, farm or landscape-level, regional-level, and so on) and hence accurate estimation of soil properties using infrared and related sensing technologies efficiently for a new prediction task. To address the comment, with context to this challenging task at hand, we will describe in more details the steps of modeling using soil spectroscopy.

For the aspects raised about Cubist, however, we are convinced that comprehending how the method works in general is sufficient. For the reader with soil science background, the crucial part represents what the rules do. This involves very basic understanding how prediction is achieved by partitioning and local regression according to simple rules, conditions and linear equations. These rules disentangle relationships in a complex data set (i.e., information-rich spectra) into simplified entities.

Cubist is commonly appreciated for the simple and interpretable rules it produces for soil spectroscopy. Nevertheless, to address the useful comment, we will make the following addition, following page 7, line 11:

"Cubist produces simple prediction rules in the form of conditions, "if-then" logical statements over the explanatory variables (i.e., spectra). These rules split the data into subsets. Here, each rule is a unique set of conditions together with the associated ordinary linear regression model. During training, these the condensed regression equations are finally made for samples in the terminal nodes. Thereby, for a particular final node, all respective split variables from it's original tree branch are potentially allowed; however, some are pruned. The smoothed regression equation with selected variables allows then to predict an individual new observation." After this statement, we will very briefly describe what committees and neighbors are (please see in the comment below).

We referred to Viscarra and Webster (2012) for a simplified summary of the statistical principles of partitioning and the meaning of the rules constructed by Cubist in context of spectroscopic modeling (page 7, lines . 12–14). To support that Cubist is a standard non-linear approach that is frequently tested in current soil spectroscopic and mapping studies — mainly due to both it's predictive and explanatory power — we will add the missing reference made to these studies (page 7, lines 9ff.). Specifically, we will cite Bui et al. (2006), Viscarra Rossel et al. (2012), Stevens et al. (2013), Pascucci et al. (2014), Miller et al. (2015), Peng et al. (2015), Viscarra Rossel et al. (2016), Dangal et al. (2019), and Padarian et al. (2019) (see section References).

RC2: More specifically, I have concerns about two points. Firstly, Cubist and RS-LOCAL were compared on a range of sites, but the latter used spiking samples from these sites while the former apparently did not, so that the comparison is questionable (of course RS-local outperformed Cubist in such conditions). And secondly, total C content was used as an example variable for both global and site-specific models, but this choice might be questionable as total C includes organic and inorganic C, which might lead to some inconsistencies (and indeed there were issues with site-specific models for total C).

AC: There are various strategies and methods of doing spiking, and many of them require 15–30 reference analyses per relatively small-area samplings performed at a particular location. For example, Wetterlind and Stenberg (2010) and Seidel et al. (2019) show the benefits of spiking samples at individual locations for national spectral libraries. Still, we found this approach too laborious, which motivated us to come up with and test an alternative strategy that substantially reduce analysis loads at plot-scale monitoring, which is innovative in our opinion. We think many studies that have dealt with global approaches but also memory-based learning have only good model performance at sufficiently high local sample density, which would be a major drawback limiting the widespread routine usage of soil spectroscopy. For RS-LOCAL, the working principle applies for groups produced in a data-driven manner (inherently considering many spectra, geographical relatedness, soil forming factors, etc.). It is important to note that RS-LOCAL requires, in addition to diverse chemical composition in a broad organomineral sense, sufficient variability in measured characteristics in the SSL. This is because variability to a certain degree needs to also be present in local feature spaces of the new prediction samples.

Our study is unique in that we adapt a transfer method that can drastically reduce the number of local samples required for systematic high-throughput monitoring purposes across many sites. This has not yet been principally done elsewhere. It is therefore possible, that spectral monitoring at a particular site is first realized with a minimal set of reference measurements. These measurements serve to measure the baseline soil conditions, while new sampling events can rely on the site-specific data and the transfer done previously in an iterative manner. We are convinced that by allowing only two training samples per site, many of them from the first campaign of Swiss soil monitoring sites, we are on the safe side.

RS-LOCAL can be considered as a very targeted selection/optimization strategy, which distinguishes it from other spiking approaches, and is a transfer learning method rather that a general regression method like Cubist. We compare strategies and not regression methods. We made sure to clearly state this throughout the manuscript, except in the abstract, where we could have differentiated a little bit more succinctly. Thus, we will improve this sentence: "Compared to general estimates of properties from Cubist, local modeling on average reduced the root mean square error of total C per site fourfold." (page 1, lines 13–14) -> "Compared to general modeling approach, the local transfer approach using two respective training samples on average reduced the root mean square error of total C per site fourfold."

Our second and third goals lay out two separate purposes and workflows (see *Introduction*). Firstly, we derived global rules for the entire collection of soils with a general optimization with unspecific adaptation. Secondly, we showed that we can achieve plot-scale adaptation of the Swiss spectral library. The second point raised by the reviewer does not change the overall

conclusion of our study: that a targeted extraction and transfer of information from the SSL makes the concept of spectrally monitoring soils at plot-scale efficient. We will address the concern made about the inconsistency induced in the specific comments section (please see e.g., response to points made: page 11, line 30; page 17, lines 12–21; page 17, lines 26–34).

RC2: Cubist procedure is poorly discussed. The advantages of RS-LOCAL are specified, but its drawbacks should also be mentioned (e.g. library subset selections for each predicted variable; need for spiking samples, which represents analytical costs; strong dependency of library subset selections on a very small number of spiking samples).
AC: Please refer to our response above. We hope that the Cubist method will be succinctly described with the proposed addition, both in terms of underlying mechanisms and its improvements made on top of M5 model trees. Please note that most of the studies have tested it for spectroscopic modeling of larger libraries. Say one can afford large samples with reference measurements for regional or site-specific adaptation purposes. As any spectroscopic approach requires validation, RS-LOCAL allows for adapting the library to the local soil conditions. Such adaptation efforts can be detected by minimal validation, that any spectroscopic estimation workflow needs, but are often overlooked.

It matters that the selected local adaptation (training) samples are representative of both the local validation and prediction (test) samples. As long as this is the case, which can be guaranteed by appropriate chemometric calibration sampling techniques, the low number of local calibration samples is actually preferred. It offers a huge advantage compared to general models that are often developed across diverse soil samples, but fail for slightly different new sites. We should not forget that the principal goals of methodologically improving and applying soil infrared spectroscopy and other sensing technologies is to make the estimation of many soil properties quicker, simpler, cheaper, and even more accurate. There are two simple means to do that: 1) maximizing information transfer from the library and/or 2) minimizing reference analyses for the target samples.

We see further room to created improved versions of RS-LOCAL. Apart from multi-response CNNs, most general machine learning and local chemometric methods need to be specifically calibrated for different soil properties. In most cases, analytical costs will be at the expense of systematic and random errors for a non-adaptive general strategy compared to a transfer with minimal analytical required such as RS-LOCAL.

RC2: I add that revision would be easier with continuous line numbering, and with all tables and figures at the end of the manuscript.
AC: The next revision will be with continuous line numbering.

RC2: I recommend major revision.
AC: Thank you.

3 Specific comments

3.1 Title

RC2: Title. Mid-infrared should probably be mentioned in the title.
AC: We thank you for the proposition and will incorporate that.

3.2 Abstract

RC2: P1L6-7. Cubist parameters (committees, neighbors) are not well known and should be briefly defined, or not mentioned. "by location grouped ten-fold cross-validation" is somewhat unclear too (but this is perhaps difficult to explain clearly).
AC: Here, we can only briefly mention how we used and trained Cubist with main hyperparameters we optimized. If it reads more clearly, we will reword as follows to avoid jargon: "ensembles of rules (committees)" instead of "committees of rules". Committees in Cubist are a form of ensemble modeling, which is widely known in statistical and process-based modeling. We will specify "nearest-neighbors used for local averaging". We will further change "by location grouped ten-fold cross-validation" to "10-fold cross-validation grouped by location", which should be clear enough.

RC2: P1L8. NABO has not been introduced; what is this?

AC: Thank you for spotting this. We will replace it with "Swiss soil monitoring network (NABO)".

RC2: P1L12-14. Relating RMSE to the range is not very informative; usually it is compared with standard deviation (=>RPD) or interquartile range (=>RPIQ). RMSE is not informative as long as distribution has not been specified.
AC: Thank you for the comment. In addition to RMSE as widely accepted general-purpose error metric, we will report the ratio of performance to interquartile distance (RPIQ) to highlight performance relative to the measured distributions. Following the arguments of Bellon-Maurel et al. (2010), we will not report RPD. It can be inappropriate for non-normal distributions. Particularly, this applies when properties such as the total carbon values are skewed as in our study; here, RPD is redundant and can even be misleading. Instead, R-squared is a much simpler and established measure of goodness-of-fit, carrying the same information content for relative comparisons.

RC2: P1L15. Dissimilarity was between subsets and validation samples, which should be specified.
AC: We thank the reviewer for this comment. The dissimilarity between SSL subsets and the respective local validation samples will be specified in the next version of the manuscript.

3.3 Introduction

RC1: [Page 2, lines 4–20] The whole first paragraph is weak, try to improve.
AC: This comment is nonspecific. We ask the reviewer to provide a clear direction towards we could work on to possibly improve.

RC2: P2L17. Reference to Dokuchaev is great! Should be done in every paper...
AC: Agreed, we are happy to have this comment.

RC2: P2L19-20. Janik et al. (1998) did not compare MIRS and NIRS, so this citation does not seem the most appropriate.
AC: Thank you. We will take your suggestion and also add a study where accuracy and discrimination of chemical soil constituents of these two energy ranges are detailed in terms of various modeling and application aspects: Viscarra Rossel et. al. (2006) [see References]. We will keep Janik et al. (1998) because the other important point we wanted to make is that there are fundamental differences in the nature and characteristics of resolving molecular groups between the mid-IR and NIR.

RC1: [Page 2, lines 21ff.] Unclear how this paragraph is related to the two goals mentioned in the next paragraph
AC: This comment is not specific enough for us to respond. Please point out what is unclear in more details then we will improve it if reasonable.

RC2: P3L15-16. Guerrero et al. proposed spiking without extra-weighting (2010) before they proposed spiking with extra-weighting (2014).
AC: Thank you for correcting our statement. We will correct to Guerrero et al. (2010) and will mention Guerrero et al. (2014) as a first reference for spiking with extra-weighting.

RC2: P3L14-29. I particularly appreciate this part. Thank you!
AC: Thank you.

RC1: [Page 3, line 24.] "These properties make it as well a transfer learning method." Sentence makes no sense.
AC: In the previous paragraph, we introduced the original concept and the research field of transfer learning. To make this connection very clear on page 3, line 24, we will modify the sentence as follows: *"Therefore, the spectrum-based learner is also considered a transfer learning method"*

RC1: [Page 3/4, lines 31ff..]

AC: The reviewer has highlighted these lines in the PDF document, but there is no comment provided. We ask the reviewer to please give some details on what can be improved here.

RC1: [Page 4, line 11] "available soils" -> available soil
AC: Thank you for detecting. The sentence will be changed to: *"The second goal was to develop general rule-based models for all available soil properties using the* CUBIST *algorithm"*

RC1: [Page 4, line 11] "a relevant subset of the remaining Swiss SSL": How was this subset chosen? How many samples were included? I see now, so add (see description below) or similar.
AC: Thank you. As suggested, we will add "(see description below)" in the revision.

RC2: P4L12. In my opinion, total carbon "content" should be specified, at least the first time; because carbon stock should also be considered. Moreover, total C is probably not the best example for illustrating the approach, as commented below.
AC: We will change the sentence to: "Further, we wanted to infer important spectral regions in the models and their chemical associations, which we illustrated with the estimation of total carbon (C) contents". Infrared spectroscopy can not directly detect the bulk density to measure the carbon stock, but this an important measure to get correct to derive carbon stocks. We suggest that we mention this briefly in the next paragraph. We eventually will modify as follows: *"For soil monitoring in general and also for determining carbon stocks, it is crucial to obtain locally accurate and precise spectral estimates of key soil properties such as organic C contents, from high soil variability of large SSLs and over time."* (page 4, lines 13–14)

RC2: P4L13. Why only "unbiased"? "Bias" has particular meaning in statistics (mean prediction error); so "unbiased" means there was no prediction error in average, which is not sufficient for ensuring accurate prediction.
AC: We agree with the suggestion made and will also carefully check this again in all parts of the updated version. We mostly use the term mean error (ME) for bias in Figures, but were not entirely consistent (Figure 3). We will stick to ME in all figures. In text, we will keep large or small bias, or will consistently use "less biased" or "more biased" when we mean that the mean error was lower or higher.

There are both bias (mean error) and variance (standard deviation of the error; SDE) components of the error, that we want to discuss, for the latter we will consistently use the term precision or adjectives more precise or imprecise for relative comparisons. We used accurate and inaccurate when we refer to RMSE in a specific context, but will now also consistently report RPIQ for relative comparisons.

RC2: P4L20. Same consideration regarding "bias".
AC: Thanks, we will make this simpler and clearer here, too: *"We therefore wanted to design a local calibration strategy using transfer learning, that would be effective in reducing (conditional) errors at monitoring plots compared to the general rules derived in the first aim."*

RC2: P4L22. Briefly recapitulating the objectives would probably be useful: (i) develop a national SSL; (ii) build general prediction models using CUBIST; (iii) build local prediction models using RS-LOCAL.
AC: We like this proposition and will sum up in this manner.

3.4 Methods

RC2: P6L1-3. Fig.1 shows some areas were not sampled. We may assume they corresponded to mountains but this should be briefly specified.
AC: Yes, we will mention that locations on the BDM grid that were not sampled were inaccessible. These were mainly points in the alpine regions.

RC2: P6L9. We may assume that total C and N were determined by dry combustion (CHN), CaCO3 by calcimetry, organic C by difference between total C and CaCO3-C, clay, silt and sand by sieving then pipette; but this should be specified. Moreover, CECpot should be defined..

AC: Thank you. We will briefly mention the methods for determination of the soil properties.

RC2: P6L18. The spectral range covered an important part of the NIR region (1333-2500 nm) so the studied spectra were not exactly mid-IR spectra.
AC: Thank you. We apologize, we have mistakenly not mentioned the selection of the mid-IR range ($4000\,\mathrm{cm}^{-1}$ to $600\,\mathrm{cm}^{-1}$) before further processing and modeling the spectra.

RC2: P6L25. Reflectance was measured and then converted into apparent absorbance according to absorbance = log10(1/reflectance). R has not been defined.
AC: Thank you for identifying. We will change as follows: *"The resulting reflectance spectra (R; background referenced) were converted to apparent absorbance (A) by* $A = \log_{10}(1/R)$*"*

RC2: P6L27. Fig.4 & 5 show spectra were not derived so "first derivative smoother" is questionable.
AC: Both Figure 4 and 5 show unprocessed absorbance spectra because the reader is more familiar with those. Hence, this facilitates interpretation. For all modeling done, preprocessed spectra were used, which we state here. We will add a brief note in the caption of Figure 4. For Figure 5, the caption should be clear enough: *"...Panel a contains the principal components subspace (PC1 and PC2) of the Savitzky-Golay first derivative mid-IR spectra, and panel b outlines the corresponding absorbance spectra, which are coloured by the total C content..."*

RC2: P7L9. "has shown excellent performance etc.": citations are expected.
AC: We fully agree. Accordingly, literature listed in the general response will be added to support the evidence made.

RC2: P7L11. I NEED SOME MORE INFORMATION ON CUBIST TO UNDERSTAND HOW IT WORKS, instead of having to look for other papers; and also to understand how you used it.
AC: Please see our response to this point made in the general comments.

RC2: P7L15. I also need to know a little bit what these committees and neighbors represent, without having to read other papers.
AC: Thank you. In line with the comment made above, we will very briefly outline these parameters and concepts in section 3.5.

RC2: P7L19. This suggests cross-validation groups were selected at random (all samples from a given site being kept together), which should be specified.
AC: The groups were not random (site), but instead the (unique) groups were randomly sampled into 10 folds for each of the five repeats. Then, the grouping information was replaced with the row IDs to derive training and test sets so that observations of the same site are kept together. We make the result of this procedure also clear on page 7, lines 16–18.

RC2: P7L20. The notion of "predefined random seeds" is not clear for me.
AC: It relates to random number generation to produce random splits for repeated cross-validation. It is important to note that cross-validation is consistent and reproducible because it is used for the comparison of approaches . To avoid technicalities, we will delete the sentence and revise this part as follows: *"The division into training and validation proportions was done in consistent and re-creatable manner using pseudo-random number generation"*

RC2: P8L6, P8L13. "equally weighted usage" and "importance measure" are unclear for me.
AC: True, we can simplify the sentence to make it even more clear. We suggest the following: "average of relative usage frequencies of a particular variable in split conditions and regressions."

RC2: P8L18. Five committees were used; but committees have not been defined.

AC: Thanks. We indeed need to mention committees in in one ore two sentences. As in the comments made above (page 1, lines 6–7), we will very briefly describe and use also less technical terminology that is more widely established: ensembles of rules (committees).

RC2: P8L19. The way calibration samples were selected is extensively described (though not always clearly); but the regression procedure should be specified earlier than P9L5 (and the way the number of PLS latent variables was determined should be specified too). It should also be specified that for each NABO site, A DIFFERENT SSL SUBSET WAS SELECTED FOR EACH PREDICTED VARIABLE, which is probably tedious.

AC: We acknowledge the second point made. Both the general scenario and the modeling particularities are important aspects. We see now that we forgot to specify that we tested between 1 and 10 PLSR components in both the RS-LOCAL selection procedure and the PLSR models made with these selections. We refered to the Lobsey et al. (2017), where the developed method is outlined in detail, but only mentioned PLSR specifically for the final localized prediction model based on the SSL samples selected and two local samples. Our next version of the manuscript will include that.

We made it clear that there 71 separate model transfers for each monitoring sites and therefore assumed that it would be clear to the reader (page 8, line 22). To make this very explicit, we will consider to modify this sentence as follows: *"We therefore conducted 71 separate sample selections from the SSL, each yielding different transfer subsets of the SSL, to test spectral-based soil monitoring using this SSL"*. We will carefully revise and lightly re-structure this section.

Yes, the subsets differ per site, this is what makes the method adaptive. It is not computational costs are evidently higher when making a transfer that is adaptive per site, but the benefits of higher accuracy out-weights the need for less chemical measurements needed.

RC2: P8L20-28. I guess this could be clarified and possibly shortened. "this SSL" L22 is unclear. Moreover, 71 models were conducted per predicted variable (only total C is considered here).

AC: We decided to focus on one property for plot-scale transfer targeting soil monitoring applications using mid-IR spectroscopy. True, we stated that we did "71 separate model transfers", but it would be indeed good if we specified *"71 separate model transfers for total carbon"* here, too. We present detailed results and can only report which analyses we did. We will replace "this SSL" with *"the Swiss mid-IR SSL presented here"*

RC2: P8L30-32. Unclear.

AC: Information leakage is the process that creates over-fit, that is when seen data of a particular independent data unit are used to predict unseen data that belong into the same data unit. Over-fit can be a result of training and validation data being not independent. We will replace information leakage with *"over-fit"* here to make the understanding more straight-forward.

RC2: P9L2-3, P9L7. Why not mi, Ki and Ni, as in P10? (but in the next paragraph i = iteration) used for validation. To avoid confusion, a specific name ("spiking samples"?) should be used for these two samples firstly used for validation SSL subset selection) then for calibration (predictions on the other samples of the NABO site considered).

AC: The reviewer is absolutely right that using subscript i here — which denotes the site number — is not adequate when we use the same abbreviation for iteratively discarding samples in individual RS-LOCAL procedures. We agree to keep the formal notation K and m for the RS-LOCAL data set and N validation samples and follow notation by Lobsey et al. (2017), that proposed the original method. In addition, we will instead of i use the subscript s, standing for the (monitoring) site. We should make the distinction by site because separate RS-LOCAL optimization (tuning) and final subset selection was done for each site. Moreover, the analysis of feature space distributions between site-specific validation samples N_s and the subsets K_s was done by site (page 10, lines 9ff.).

RC2: P9L4. "model performance (RMSE) for the two local calibration samples": here these two samples were used for validation. To avoid confusion, a specific name ("spiking samples"?) should be used for these two samples firstly used for validation SSL subset selection) then for calibration (predictions on the other samples of the NABO site considered).

AC: We have a strong preference for calling m samples *"local calibration samples"* (page 9, lines 2–3). The role these samples have in RS-LOCAL differs from a typical spiking approach (e.g., without or with extra-weighting). Once the final selection of

RS-LOCAL training data from the library is done, this set is augmented with the m set to perform final PLSR calibration to predict the local or plot-specific validation samples. In Lobsey et al. (2017), m samples are called site-specific samples; here, we have a more complex scenario. This has lead us to adapt this notation because individual monitoring sites consist of a small-area plot ($10\,\mathrm{m} \times 10\,\mathrm{m}$), for each of which a separate transfer is done. The two respective plot-specific calibration samples have analytical values that are specifically determined for making re-sampling based optimized SSL sample selection procedure. Calling them *"site-specific samples"* in the context of our work would not evidently discriminate them from the validation samples from the same area, represented by the third to the last observation of the time series measured. Alternatively, we could also name them local tuning samples because of their role in the SSL the selection

RC2: P10L3-4. "the two local validation samples were excluded": cf. my previous comment! =>spiking samples"? But the important point is: WERE THE TWO SPIKING SAMPLES PER NABO SITES USED BY CUBIST for this comparison? APPARENTLY NOT, WHICH IS VERY QUESTIONABLE, because RS-local used them so got local information; and in such conditions, of course it will outperform Cubist. ACTUALLY SUCH COMPARISON DOES NOT SEEM RELEVANT. Comparison would only be relevant if RS-local did not use two spiking samples per NABO site, or if Cubist was run in a way that allowed spiking.
AC: . Here we refer to our response made in the general comments sections. There are similar specific points made that pointed in this direction: e.g.,

RC2: P10L9-10. Ki, mi and Ni were Ksite,i, msite,i and Nsite,i P9.
AC: Thank you for pointing out this inconsistency. We refer to the answer given in comment on page 9, lines 2–3, and will correct accordingly.

RC2: P10L11-12. Unclear. Distance calculation does not seem to consider Ni samples (which were however mentioned L10). And I don't understand these two distributions of distance. Moreover, in order to quantify the similarity between Ki and mi (+Ni?) samples, why not considering DISTANCES BETWEEN THEM, instead of distances to the center?
AC: Thank you. We agree this is contradictory because we effectively compared distribution of N_{site} validation samples and K_{site} subsets (note the new notation *"site"*). We will correct this mistake. The reason behind the second aspect mentioned is simple: by referencing all points to the robustly calculated center of all data and reporting those distributions, we solve the problem of calculating and aggregating full-factorial combinations of distances between all validation points and points of the selected subsets. Instead, we can simply compare the two distributions of distances to the center ($\{N_{\mathrm{site}}; K_{\mathrm{site}}\}$) for each site, without loosing any information (i.e., Figure 7).

RC2: P10L13-21. The Mahalanobis distance is well known in spectral analyses (much more than Cubist, described in 10 words P7L11...) so this part could easily be condensed
AC: We agree with that. However, we do not detail the well-known general Mahalanobis distance method. Instead, we outline the general procedure of distance calculation mentioned in the response above, and refer to a robust method that we used for calculating multivariate location and covariance estimates plugged into the distance equation: the minimum covariance determinant (MCD) (we leave the details to Hubert and Debruyne (2010) for the interested reader: *"Note that the MCD estimator can only be computed when $h > p$ [h := number of observations; p := number of dimensions], otherwise the covariance matrix of any h-subset will be singular. "*). Unfortunately, the MCD estimator has not (yet) found prominent application in soil spectroscopy apart from robust outlier detection using this robust Mahalanobis distance method for robust outlier detection (see e.g., Filzmoser et al. (2005))

RC2: P10L14. distance to nearest neighbor becomes similar to distance to farthest neighbor? Well... surprising; and probably useless here (the Mahalanobis distance does not have to be justified).
AC: That is the interesting aspect of the curse of dimensionality. We will reconsider to remove.

3.5 Results

RC1: [Page 11, lines 3–10] Overall badly written, improve.
AC: It will be difficult to address this comment as it is not clear what could be improved, however we will carefully revise and check if we can improve anything. It would be useful to know what aspects should be removed (content, phrasing, grammar).

RC2: P11L11. RPD AND/OR RPIQ SHOULD BE SPECIFIED, OTHERWISE PREDICTION ACCURACY CAN HARDLY BE ASSESSED (R^2 expresses proportionality, not similarity; and RMSE is poorly informative when not compared to distribution parameters; range is not sufficient).
AC: We agree with the reviewer, we should certainly have also reported RMSE and RPIQ here. In line with the RPIQ presented for the plot-scale transfer with the general approach (Figure 6), we will add RPIQ in Table 3 for the CUBIST results.

RC2: P11L13. What humus represents has not been specified in section 2.2. Moreover, R^2=0.87 for SOC, so not >0.9.
AC: This is true. We used humus but the correct term is organic matter. We propose to remove this property because it is based on multiplying the organic carbon determined with the Walkley-Black method with a constant factor of 1.72, which is a linear transformation. We admit that there is no added value in reporting organic matter. This factor was established for arable soils, and generally soil functional group composition of carbon compounds can differ considerably. This particularly applies for organic soils, where some soil even exceeded 100% organic matter due to erroneous extrapolation of this factor.

RC2: P11L20. Skewness has not been presented so we don't know which variables were skewed.
AC: Thank you. Following the comment made in page 12, Table 1, we will report skewness instead of the coefficient of variation in Table 1. Coefficient of variation is indeed evident from the mean and standard deviation of the measurements. in Table 1. We further depicted the kernel density distribution for all modeled properties on page 13. Skewness can be deduced from Figure 3. Thus, we will make a reference to both Table 1 and Figure 1, which was missing: *"We found marginal local bias at the largest values, mostly for variables with positively skewed distributions such as total C."*

RC2: P11L22. Organic matter has not been presented in section 2.2. How was it determined?
AC: This is true, we should have done this. We will address this in section 2.2 as explained in the earlier comment made.

RC2: P11L25-28. This is confusing: for total C content, 0.834%, achieved with 209 bands, was not the lowest error, as lower errors were achieved with 105 and 90 bands; while RMSE=0.84% in Tab.3. And I do not understand what test RMSE (RMSEtest, also used in Fig.4) represents (vs. RMSE before).
AC: We thank for the observation and apologize for this erroneous statement made from our side. We will correct accordingly. As the reviewer correctly remarks, we are slightly inconsistent because we name the error just root mean square error (RMSE) (Caption of Figure 4) and then proceed with test RMSE.

RC2: P11L30. CONSIDERING TOTAL C IS PROBABLY NOT THE BEST EXAMPLE as it includes organic and inorganic C, which involve very different contributing bands. I would recommend organic C as example.
AC: Although organic carbon has a more distinct role than inorganic carbon in soil functioning, we do not expect that the main message changes. We had methodological focus, and the aims could more efficiently addressed with more variation and measurements available. for total C. Because we initially had more data on total carbon than organic carbon, both the general approach and the local approach were detailed with mid-IR spectroscopy using the same soil property. Organic C can be expected to follow the trend of total C at a particular site.

Moreover, Helfenstein et al. (2021) support our finding of diverse selection done on plot-level, for the regional level investigated there (wide geographic selection), which similarly was found by Lobsey et al. (2017) (personal communication, Raphael Viscarra Rossel).

Ultimately, selection of this particular property does not change our message: that in most cases plot-specific transfer from the SSL yields accuracies that have potential for monitoring with minimal local reference analytical knowledge, which can discriminate a realistic range of changes that occur over decades. One important message is that his transfer concept can be

expanded to other monitoring sites, where more quantitative characteristics can be collected, which gains statistical power for detecting soil changes over more and more representative areas of Switzerland.

. .

RC2: P12Tab1. Instead of CV, which can be calculated easily, I would prefer skewness. Moreover, what humus represents should be specified. Negative concentrations should be avoided (and replaced by 0). DNA has not been introduced in the presentation of methods. Why is CaCO3 not mentioned for NABO? (while it is in Tab.5 so was determined on these samples)
AC: Thank you for discovering these inconsistencies. We will remove the all parts that treat DNA. The estimation is too generic and the property is not well predicted. However, we forgot to mention $CaCO_3$, whose variation was explained rather well by mid-IR modeling; hence, we should have listed (also on page 20, lines 4–5).

RC1: [Page 13, Figure 3, Panel "Organic Matter"] How was OM determined? Just a factor and based on TOC? Is it the same as humus above? How is >100% possible?
AC: We described the method for OM analysis in the method section. Also see the comment above.

RC2: P13Fig.3. Curves suggest R^2 was not calculated from linear regression (as mentioned P7L30). RPD and/or RPIQ should be specified. mg L-1 for extractable elements? (mg kg-1 in Tab.1)
AC: We disagree the first comment made. We exactly calculated as stated in section 2.5.1 (page 7, line 30). R-squared is correctly calculated, but the trend line shows a spline approximation to visualize better the trends in the residuals. We will add a short note to the caption.

RC2: P14Fig.4. Figure caption should present parameters that are important to understand the figure, but does not have to fully present (and justify) the methodology. Most contributing bands included bands that have either been assigned to organic compounds or carbonates, which suggests choosing total C as example was not necessarily appropriate.
AC: We thank for outlining how we can further simplify the caption, which we will change accordingly in the next version. For the second point, we do not follow why this should be different for the model of the organic carbon content, and hence "inappropriate". There is fewer rules, yes, but soils with relatively high organic carbon contents can have low or high carbonate peaks (peak height is relatively linear to concentration). These are interaction effects that the model must inherently account for in prediction.

RC1: [Page 15, lines 5–10] "...with 55 samples from the SSL (K), 10 sampling events (B of size K at each iteration, and 10% reduction (r) at each iteration (Figure 5). This effectively yielded 52 transfer samples...] I don't get this, why dont you just use the 55 samples and add the 2 spiking samples?
AC: Thank you for noticing. We will correct to: *"...with 55 samples from the SSL (K), 10 sampling events (B* of size K at each iteration, and 10% reduction (r) at each iteration (Figure 5). These 55 transfer samples from the SSL were combined with the two site calibration samples previously used to supervise the selection from the data source, ...". We here detail the tuning results of the example site to show more context.

RC1: [Page 15, Table 3] Where is Table 2?
AC: We apologize, this is a LaTeX typesetting mistake while preparing the pre-print. We will correct the table numbering in the revised manuscript.

RC2: P15Tab.3. "to achieve many test-train data combinations and provide... generalization" is useless here. RPD AND/OR RPIQ HAVE TO BE SPECIFIED. Moreover, % is not a SI unit and should be avoided (also for avoiding confusion: e.g. SOM was 5% initially and increased by 10% means it reached 15% or 5.5%?), and replaced e.g. by g kg-1, or possibly g 100 g-1 for clay, silt and sand. No decimal for humus?
AC: Thanks! We agree that this note is not necessary in the caption because we have explained in the methods section. We agree and will consistently report carbon contents in $\mathrm{g\,kg^{-1}}$ soil and particle size fractions in $\mathrm{g\ 100\,g^{-1}}$ soil in the entire manuscript as noted earlier.

RC2: P15L5. K? Ki? k? I understand K=55 (or K=52? or K=54=52+2?), while P9L14 mentions K=50 and Fig.5 both 50 and 55, WHICH IS CONFUSING (P9L9-10: "Parameter k is both the number of samples drawn from the (...) library (...) and the number of samples of the returned SSL subset"). Moreover, I understand this optimization of K/k, B and r was for total C, but was possibly different for other variables, which should be specified.

AC: We made a wrong statement here, thank you for reporting. Will will correct this. Yes, parameter k of RS-LOCAL is both the number of SSL samples randomly selected in the re-sampling step, and also the target number of SSL samples returned by the algorithm (please see also Lobsey et al., 2017).

RC2: P16Fig.5. Spectra are not first derivatives as erroneously mentioned in the caption (smoothed spectra?). "This subset was most accurate" does not seem appropriate (prediction was most accurate using this subset). The fact that PC1 and PC2 represented <40% of total variance is surprising, I've never seen such small value; I guess SNV spectra should have been used rather than smoothed spectra. And it should probably be reminded that PCA was built using all 4374 spectra (if I've well understood).

AC: Thanks for spotting this! We disagree that we should have used spectra transformed with standard normal variate. We were a bit less surprised about this, but we consider mentioning this interesting fact. We will give some references to comparable mid-IR SSLs and will give possible explanations. For correct interpretation, however, we need to use the spectra that we have effectively used for modeling (Savitzky-Golay derivative preprocessing with a window size of 35 points and third degree polynomials.)

RC2: P17L5. R^2 does not deserve much attention (it is useful mainly in the abstract section, for readers that are not familiar with spectral analyses, and also in tables or figures).

AC: We thank for the aspect mentioned. A single accuracy measure can not describe all aspects of model interpretation and comparison. In Figure 6 we also show R^2, the mean error and RPIQ, hence we also think that it is useful to mention R^2 here in the text, too.

RC2: P17L1-7. RMSE, R^2, RPD and RPIQ SHOULD ALSO BE SPECIFIED FOR THE SET THAT INCLUDED ALL VALIDATION NABO SITES, instead of mean RMSE and bias (which are questionable). If the two spiking samples per NABO sites were not used for Cubist, better predictions with RS-local than with Cubist is not surprising, and actually, the comparison does not seem relevant (cf. my comment regarding P10L3-4).

AC: We agree with the reviewer that we should also report average conditional relative errors by site. Nevertheless, absolute errors and their contributions are important. Some of the sites have only minimal fluctuations, but the RMSE reported might be sufficient to discriminate changes if any significant ones occur. We give absolute bias to account for possible zeroing out by averaging. Our next version will also report R^2 and RPIQ consistently.

RC2: P17L10. RPIQ=3.08: is that again an average?

AC: Yes, thank you. We will state that more explicitly in the next version.

RC1: [Page 17, Lines 16–17] "The two local calibration samples and the 12 validation samples on the upper right corner were close to each other in the PC1-PC2 subspace (Figure 5, panel *a*, *left* and *right*; range PC1: 9.2 to 11.0; range PC2: 4.9 to 7.5)" -> I would actually observe that especially the validation samples are at the edge of the PC1PC2 space which seems not ideal to me.

AC: We perform a contextual validation of the scenario, Figure 5 shows the example of one site. If we did a simple calibration split—-as it is commonly done for simple isolated model development and evaluation across all data—such point could be valid (validation are then typically chosen by calibration sampling of spectra or multiple random splits are generated); however, here a transfer is validated for a single site. In fact, the majority of the depicted principal component space is formed by the transfer set. To be realistic, the scenario assumes that a new carbon monitoring site is initiated at once while a library is already present; this would anyway require baseline estimates acquired from measurements.

RC2: P17L12-21. THIS "DISSIMILARITY" QUESTION MIGHT BE DUE TO THE PARTICULARITY OF TOTAL C, which includes organic and inorganic C. Indeed, Fig.5 shows that samples from the validation site (65COR) were carbonate-free (no peak at 2500cm-1), while some SSL samples selected for calibration contained carbonates (peak at 2500cm-1). So it

might be assumed that in the SSL subset selection procedure, some carbonated samples were useful for reaching appropriate total C prediction on the two spiking samples; and to some extent, this might be considered a kind of overfitting (i.e. unsuitable selection of carbonated samples for minimizing RMSE on the two 65COR spiking samples). To dissipate such doubts, RS-LOCAL RESULTS SHOULD BE PRESENTED FOR ANOTHER VARIABLE, E.G. ORGANIC C.

AC: It is true that RS-LOCAL is rather unusual compared to most of other methods that put local constraints, because it is assumption-free and completely data-driven. This gives the method the flexibility to deliver less biased and more accurate site-specific predictions at minimal analytical efforts without any assumptions about soil type, geography, depth or spectral similarity. Many local methods capable of reducing larger libraries are good at reducing slope trends and the random error term in predicted vs. measured, and thereby improve the bias. However, these methods still need a relatively high amount of local reference measurements (often more than 15), while purely local calibrations have been successfully established with as few as 30 samples.

We do not support the point made that we need results on another property such as organic carbon. The carbonyl group has a very isolated peak in the spectrum (2522 cm^{-1}). We focused on total C due to more measurements, which strengthened the point made about yielding maximal variability and showing the efficacy of the targeted selection and transfer. Carbonate content was clearly not the main driver in the first two principal components, which is further supported by Helfenstein et al. 2021 where PC1 also discriminated total C contents but where only a few soil samples had carbonate peaks in the spectra. Table 5 supports the diversity of selection in many other soil properties, and interestingly the standard deviation of the absolute difference. We may consider to move Table 5 to the appendix.

Unfortunately, we can not tackle local feature importance since it too much to address in one study, but think this should be done in following work. However, we previously made an in-depth exploratory analysis of which original unprocessed spectra were selected in the individual rules derived from Cubist using preprocessed spectra. Here, to our surprise, we found that the subsets that each rule in the first ensemble typically selected contained heterogeneous mixture of characteristic patterns of soil spectral features. For example, some of the spectra in the groups had no carbonate peaks and others had gradually differing carbonate peak heights. Still, predictions of total carbon were relatively accurate. Although Cubist works differently than selection and PLSR modeling leveraged by RS-LOCAL, it clearly showed that carbonated soils were typically used to make regressions for the carbonate-free soils in a general approach that lacked site-specific calibration samples. Such inherently local but at first sight less intuitive mechanisms of dissimilarity seem to occur in both general and site-specific local approaches, independently of local transfer. However, such mechanisms are leveraged also in random forest (bagging) and methods that leverage boosting or similar mechanisms (e.g., Cubist).

Meanwhile, using the present Swiss mid-IR library, Helfenstein et al. (2021; accepted in SOIL) show a regional application of RS-LOCAL for diverse soil samples from two drained peatland regions where only a few mineral and organic soils with lower total carbon contents had characteristic carbonate double peaks (RPIQ =). Still, the predictions for soil horizons that were that carbonate-free and aligned

When RS-LOCAL was first presented, one of the indications were that a larger spectral library to some degree needs to contain sufficiently high organic carbon contents comparable to the local set. An assumption there was that the mineral matrix alone is sometimes less informative to derive data-driven selection and a consistently accurate local adaptation for the entire range of measured values. In our scenario, we also discriminate validation results per site and aggregated over time (Figure 6). This shows that improvement in RMSE for many of those locations that were improved by the local adaption strategy with two additional training samples goes together with an increase in the R^2 (linearity inpredicted vs. observed) and/or reduction of bias (slope or offset of the model assessment regression).

Regardless of these points made, global modeling approaches are still prone to deliver locally biased estimates from spectra (see e.g.), which is unspecific model (over-)fit due to a lack of transfer. We believe this is often neglected, although such effects have been repeatedly reported when applying regional, national and more extensive libraries to new locations. Nevertheless, also local adaptation with RS-LOCAL needs — as any other statistical method – critical assessment with validation metrics and an adequately representative calibration sampling scheme (chemometrics) to select site-specific samples with analytical measurements needed.

RC2: P17L26-34. The important point is that samples used for calibration cover the diversity of validation samples mineralogically and texturally; and if possible have larger maximum Y and lower minimum Y (Y being the predicted variable); in my opinion, this is more important than similar Y distribution for calibration and validation samples.
AC: True, the described characteristics relate to major aspects of the chemical composition of soils. These aspects make the main discrimination of the spectral feature space of the respective selected subsets of the SSL, which is certainly expected. The point we have made here is that samples with highly variable soil properties related to chemical composition are selected from the library for each location. Part of it is indirect (reverse) causation because we reasonably assume that were high differences in pH and CEC etc. go in hand with soil compositional variability, however does not contract the point made by the reviewer.

RC2: P17L32. What ">6.8%" refers to is unclear.
AC: We will correct as follows: *"Further, the measured clay and $CaCO_3$ contents were markedly different between the* RS-LOCAL *selection and the local validation sets (mean absolute median differences of 8.5% clay and 8.9% $CaCO_3$)"*

RC2: P19Fig.7. As previously mentioned, it would be more useful/relevant to study Mahalanobis distances between samples from each NABO site and corresponding Ki samples (cf. my comment regarding P10L11-12). I'm very surprised by Mahalanobis distances >20 (3 is often considered a threshold for spectral outliers); this suggests there was some issue in distance calculation. Moreover, in either panel I do not see the site-specific samples of some NABO sites (especially the first ones).
AC: Yes, we did study the Mahalanobis distances between samples from each NABO site and corresponding K_{site} samples. We refer to the explanation given in the response to the comment of page 10, lines 11–12. We see no issue in the distance calculation, we have correctly implemented the robust Mahalanobis variant using the MCD estimator (via the square root of `stats::mahalanobis()` using in plug-in estimates of the MCD center and covariance terms calculated with `robustbase::covMcd()` in R on first 10 principal components of the preprocessed spectra.). Nevertheless, we will check all calculations made. While going through this section again, we have noticed that we should have also cited earlier literature of Filzmoser et al. (page 10, line 15), which we will do in the next version of the manuscript.

RC1: [Page 17, line 30] "the SSL" -> from the?
AC: We will change as suggested.

RC1: [Page 17, line 31] "...for measured C and OC." -> for measured C and OC, *respectively.*
AC: Thank you, we will change this.

RC1: [Page 17, line 31] "...the measured clay and CaCO3 was markedly different" -> were
AC: Thank you for correcting.

RC2: P17L13. The fact that PC1 and PC2 represented less than 40% of total variance is surprising.
AC: We were less surprised about this. We kindly refer to our response made to the comment on page 16, Figure 5.

P17L19. The notation Krs-local has not been used yet.
RC1: [Page 12, Table 1: Humus [%]; column "Max"] What is humus? OC is above so what is this. And how can it be up to 115%?
AC: Thank you for the comment. We will change the wrong term "humus" to *soil organic matter*, and will describe how it was determined in the method section, accordingly. The incorrect and large values are due to an artefact: the conversion from organic carbon contents to organic matter contents is done by using multiplying the Walkley-Black measurements with a constant multiplication factor of 1.72, which was established for agricultural soils with organic carbon contents below 50 g kg^{-1} soil. For some organic soils, functional composition is fundamentally different so that uncertainties are too high and the organic matter contents are wrong/unrealistic ($< 100\%$); we will delete those value). We are considering to remove the model reported for this characteristic, because it is not always applicable and because the response is a simple linear transformation from the total organic carbon content.

RC1: [Page 18, Figure 6] These are important results for the comparison, unclear for me is if the methods differ in predicting the trend over time, which seems more important for monitoring than getting the actual value right

AC: From our perspective, all four metrics presented show first evidence to show to which degree one can potentially predict trends over time for such a plot-scale transfer. We agree with the reviewer that establishing trends is of utmost importance for detecting and interpreting effects of a changing environment and human interventions (i.e., agricultural management) on soil characteristics.

We are actually considering to also report the slope of the regression of predicted vs. observed values per site, moreover directly in relation to bias (bias vs. slope for individual sites for both global modeling and adaptive plot-scale transfer). The latter would allow to see whether improvement in the bias can be attributed to slope and/or offset correction. We would like too keep our focus on both bias and variance aspects, however we agree that establishing a direction of changes with more but less precise measurements is the key point of the spectral sensing paradigm.

To sum up, this study lays a first basis that spectroscopic diagnostics can be useful and (sufficiently) accurate for such plot-scale soil monitoring. Further steps need to be made in following work, because it's simply too much to cover so many aspects in a single paper.

RC1: [Page 19, Figure 7] **AC:** Reviewer 1 has marked the caption of the figure, however there seems no comment made for the highlighted lines in the PDF. We spotted a typo: "individual tranfer modeling" -> Was it this correction or is there something else unclear?

AC: Thank you. We will correct to "transfer".

RC1: [Page 20, Table 5, caption] "the across 71" -> delete [*the*]

AC: Thank you. We will correct it.

3.6 Discussion

RC1: [Page 20, Lines 7–8] "In agriculture, both measures are key factors for fertilization recommendations." -> I dont really think so, important predictions would be for availble N, P and K. And pH is important to determine liming requirements.

AC: We will provide some references related to the effects of CEC and pH on soil status and fertilizer management, which we missed to cite. The reviewer seems to think that CEC is not necessarily a key control on fertilization recommendations. If so, it would be good to get us some references to back up the argument. We agree that pH has importance for liming, but besides it a major factor that directly controls the plant-availability of key elements. This is the main reason why liming is done in the first place. We appreciate the insightful suggestion that available N, P, K are key controlling factors for yields and quality; however, we intended to discuss possible implications for applications with those spectroscopic estimates that had reasonable performance in the library. Knowledge of agronomic importance of key soil properties can be assumed from the reader. Therefore, we hope that the reviewer agrees with the decision made to only discuss sufficiently accurate modeling results with direct links to environmentally and agronomically relevant soil characteristics and functions.

RC1: [Page 20, Line 14] "(Stenberg and Rossel (2010))" -> delete one

AC: We will remove the extra bracket.

RC1: [Page 20, Lines 14] isnt that obvious, it is organic and inorganic C?

AC: We agree that the model importance confirms what is expected from a soil chemical perspective, however it seems relevant to mention the different complexity mapped by the models' prediction mechanisms. Following the comment, we will make it evident that this was expected.

RC1: [Page 20, lines 15–16] "...we had about four times for training data than OC..." -> Why is that? Do not all soil samples have SOC?

AC: The majority of BDM samples have now also organic carbon derived from dry combustion followed by subtraction of carbonates while the presented data had only organic C measured with the Walkley-Black procedure. Accordingly, the

Cubist model and results will be updated for the next version of our manuscript. We do not expect substantial changes in the interpretation, but eventually improved performance of the global model.

RC1: [Page 23, line 2] "we tested RS-LOCAL Lobsey et al. (2017)" -> by
AC: Thanks for spotting the typo.

RC1: [Page 23, line 6] "An advantage of the method is that it can further deal with" -> can deal better?
AC: We agree. This is a good suggestion and we will change accordingly.

RC1: [Page 23, line 8] "convolutional neural networks" -> (CNN)
AC: Thank you for spotting the abbreviation defined earlier. Along this line, we will also abbreviate "deep convolutional neural network" with "deep CNN" (page 3, line 26).

RC1: [Page 24, lines 3–4] "Nevertheless, thee sites of the NABO long-term monitoring program has not undergone substantial changes in OC (Gubler et al., 2019) -> That is not what I see in Fig 2. May be have a closer look at arable sites?
AC: Figure 2 only is for illustrating the local modeling scenario we made. It shows changes and modeling results only from conceptional purpose. Gubler et al. (2019) report only minor changes for the ensemble of monitoring sites, but point out that individual sites showed both increasing and decreasing trends per decade (-11 to +16% relative change per decade ("4 sites with declining trends, 9 sites with increasing trends, and 9 sites with stable or indistinct trends"). We could briefly discuss the implications of these findings and link them to the metrics of plot-scale transfer.)

RC1: [Page 24, lines 10–13] the current SSL includes soils that contain between 0.1 and 58.3% OC (Table 1). Since organic soils can have up to 50% OC, organic soils might be underrepresented in the current SSL. For this reason, we also tested the present Swiss SSL with a case study and augmented it to better represent organic soils, using new soils from two peat land regions in Switzerland (Helfenstein et al., 2020, submitted)." -> It seems strange to induce here new results from another paper, better delete.
AC Thank you. We understand the concern made. When we submitted this manuscript, we should have very briefly recapitulated the specific findings made by Helfenstein et al. (2021), where this national mid-IR library was evaluated further in this aspect of representativeness. However, we disagree to delete because we think that the referenced paper, meanwhile accepted for publication, strongly supports the point we made about the relatively low representation of organic soils in the current Swiss mid-infrared library. Still, the approach of regional prediction by Helfenstein et al. (2021), using as few as 5 additional analytical reference measurements made for the regional set in addition to the library, predicted total carbon contents of very diverse soil samples with a range of mineral and peat soils quite accurately (RPIQ = 4.93 to 7.09, RMSE < 3.2%).

For organic soils (> 10% organic C), there was comparably lower predictability (RMSE = 2–3% total carbon content) than for mineral soils (RMSE < 1%). This is likely attributed to compositional complexity of organic soils but more importantly seems also liked to under-representation of organic soils (< 2%) in the current library. We will make sure to highlight this in the next version of the manuscript.

RC2: P20 Tab.5. Similar Y distribution for calibration and validation samples is not the most important point (cf. my comment regarding P17L26-34); so the interest of this table is questionable. And table's title should be checked.
AC: We thank for mentioning the typo, which we will fix. Table 5 contrasts your general statement made: the percentiles of the respective selected SSL subsets are markedly different from those in the local samples for six key soil properties. The dissimilarity in the spectral feature space of the K_{site} site-specific samples (Figure 7; new abbreviation) is effectively mapped to substantial chemical diversity in terms of properties, while the spectral dissimilarity in strict sense typically relates to soil chemical composition. That means that diversity in the spectral space is confirmed by the chemical reference measurements of the selections made with respect to the site-specific conditions. This gives a convincing (broad-scale) picture of the mechanism(s) of (instance-based) transfer of to the reader: that plot-scale transfer (inherently) yielded dissimilarity. To give an example, mean absolute differences of the median clay contents between the 71 RS-LOCAL selected subsets from the library and the corresponding local (validation) samples was 8.5%, while it was 2.3% for total organic carbon, which signifies considerable differences in the transfer distributions.

RC2: P20L2. I would like some interpretation/discussion of parameter optimization (100 committees, 9 neighbors), when compared with other works that have used Cubist, in particular at comparable scale. I would also like some discussion about bias at largest values (cf. P11L20), probably due to small numbers of (calibration) samples with such values.
AC: True, we like both wishes. We can briefly touch upon these aspects in a linking paragraph. In fact, bias at largest values, often found at the tails of tree or tree-derived models (rules), is a sample representation issue. This evident but necessary conclusion further evidence for the recommendation made in the section "Future applications and updates of the SSL" (page 24, lines 10–14). Regarding the tuning decisions made, we found significant performance improvements when increasing from 20 to 50 and 100 committees. Other soil spectroscopic studies have used less committees, we guess mostly for computational reasons.

RC2: P20L10. "mostly comparable" is optimistic (RMSE=0.2-0.4 vs. 1.2%); but experimental conditions were different (both cited papers used representative calibration samples for achieving the cited results). Moreover, RMSE IS NOT APPROPRIATE FOR COMPARING PREDICTIONS WITHIN SETS THAT HAVE DIFFERENT DISTRIBUTIONS; RPD and RPIQ would be much preferable (anyway, even using RPIQ, comparison between studies is often difficult due to differences in sample set diversities, calibration sample proportion and representativeness, etc.).
AC: We agree and we will revise this statement, reporting in addition RPIQ and R^2.

RC2: P20L13. It can hardly be assumed that mineralogical diversity and variable ranges were larger over Switzerland than over France or over sub-Saharan Africa (i.e. the cited papers).
AC: This part indeed needs some clarification. We do not make a direct link to soil compositional diversity here, but rather make an accuracy comparision: *"The accuracy of our estimates for the properties that have direct chemical links, through compound-associated absorptions,were mostly comparable to established continental or country-specific mid-IR SSLs. For example, Clairotte et al. 2016) achieved RMSE= 0.2% for OC using mid-IR and the spectrum-based learner for local predictions, while Sila et al. (2016) reported RMSE = 0.4%."*.

RC2: P20L17. RMSE is not appropriate for comparing two variables.
AC: Thank you, we will also provide RPIQ and R^2.

RC2: P21L1. I wonder if assignations reported for Australian Vertisols are suitable for Switzerland.
AC: Organic matter, mostly derived from plant material and transformed by microorganisms, is complex in its chemical structure, stabilisation mechanisms and association to different particle size fractions. However, the fundamental mechanisms to predict carbon contents by characteristic C-H absorptions in alkyl C compounds and other functional groups of organic carbon are suitable because the absorptions in this mid-IR region are distinct and have relatively low overlap with other compounds. This applies regardless of soil type and origin. The attribution and the relative importance is the only aspect we imply here.

RC2: P21L5. The fact that bands assigned to quartz contributed to C prediction should be discussed.
AC: Thanks for the suggestion! We will briefly discuss. This is frequently reported in literature. Actually, this shows the potential of soil chemical diagnostics that is leveraged by the holistic fusion of infrared spectroscopy and statistical learning in pedology context.

RC2: P21L16-17. For some variables Cubist prediction accuracy was suitable for most users' needs (e.g. RMSE=5% for clay or 0.3 for pH), but this was less clear for others, e.g. the possible use of SOC prediction model with RMSE=12g kg-1 SHOULD BE DISCUSSED.
AC: Any evaluation metric and diagnostic model outcome has to be judged in light of the application its contextual requirements. We will revise this section based on the comment made. A SOC estimation with RMSE = 12 g kg^{-1} is not useful for local monitoring of agricultural soils over a time scale of decades, but can useful for characterizing soil variability in a ecological landscape context (mineral to organic soils).

RC2: P21L25. RMSE over all NABO validation samples would be more appropriate than the average of RMSEs per NABO site.
AC: You are absolutely right. We will report both conditional and aggregated evaluation metrics for plot-scale transfer.

RC2: P21L30-P22L2. Again, this might be specific to subsets selected for predicting total C, and conclusions might be different if examining subsets selected for predicting e.g. organic C.
AC: For the soil properties that we modeled, we have reached the same conclusions. We refer to the answers we have previously given.

RC2: P22L4-9. Such hypothesis is questionable. Indeed, several works have demonstrated that predictions at local scale are more accurate in general when using (enough) local samples only than when using libraries that cover larger areas, even when these libraries are spiked with some local samples and/or when suitable library subsets are selected (e.g. Wetterlind & Steinberg 2010 doi: 10.1111/j.1365-2389.2010.01283.x; Gogé et al. 2014 doi: 10.1016/j.geoderma.2013.07.016; Guerrero et al. 2014; Guy et al. 2015 doi: 10.4141/CJSS-2015-004; Lobsey et al., 2017; Seidel et al. 2019 doi: 10.1016/j.geoderma.2019.07.014; etc. however better Random Forest predictions were achieved when spiking a regional library than with local samples only by Nawar and Mouazen 2019 doi: 10.1016/j.still.2019.03.006). And this is contradicted by P22L22-25.
AC: . The references mentioned by the reviewer support that the hypothesis is reasonable regarding local model calibration. The statement on P22 L22–25 is not contradicting it, but points out these requirements on the spectral libraries.

There is nothing wrong with the similarity mechanisms proposed at local scale in context of the articles listed by the reviewer. These are indeed well established and are successful under the conditions of sufficient sampling densities over relatively small areas and sufficient local soil variability. We here first talk about diverse calibration sets, as they are selected via instance-based transfer using RS-LOCAL, which is evident from the previous paragraph and the following sentences. To make that more clear, we will specifically say (page 22, lines 3–5): *"Nevertheless, we can yet only speculate about how and why such diverse calibration sets that are achieved through* RS-LOCAL *transfer are able to leverage accurate local calibrations."*.

Our data and results shows consistent evidence that dissimilarity and diversity in soils (see also Helfenstein et al., 2021) further means for achieving local transfer (e.g., subset selection (instance-based) and/or feature-based transfer). Rather than restricting ourselves to similarity-only concepts, we believe we should make use of both concepts—data-driven and distance based selection and transfer—and take best of both worlds. This is a hypothesis, but given our experience with the methods and the fact that both frameworks have their validity, motivates to further advance the methods to increase the accuracy, computational efficiency and reduce analytical measurements needed significantly.

RC2: P22L6. Drawbacks of RS-LOCAL should also be considered, e.g. SSL subset selection has to be run for each variable.
AC: "Following the requirements mentioned in the introduction, a well-established and sufficiently diverse larger-scale soil spectral library needs to be available to make regionally local spectroscopic model estimation." We will add this point to the discussion made on page 22, line 55.

RC2: P23L5-6. But RS-local requires observations on local samples (two spiking samples per NABO site here), which represents a cost (i.e. the cost of spiking); while e.g. SBL does not. Moreover, much "responsibility" is put on these spiking samples, which have to be characterized perfectly because their spectra and conventional analytical data determine SSL subset selection; moreover they should represent the local site (which might be questionable if, for instance, soil properties are affected by soil management after t0).
AC: We appreciate the comment made. To be fair, we would like to incorporate these perspectives, which we will do concisely in the next version submitted. In our opinion, reliance on a few local samples needs to be carefully done and monitored in sophisticated manner, using management. The good thing here is that mid-infrared spectroscopy is quite reactive to compositional changes such as liming or changes in the relative composition of soil organic carbon fractions.

RC2: P23L23. In my opinion this section is too long; it should propose some perspectives, but not discuss them extensively. What was done is more important that what could be done.
AC We thank for the advice. This section is devoted to have some meaningful prospects so that we focus on what could be done. However, we will go over the section carefully and check where we make it concise.

RC2: P23L29-30. Similarity criteria would select library subsets very different from those selected by RS-local, at least regarding total C prediction; so rather than improving RS-local as mentioned L25, it seems this would be a different procedure.
AC We disagree. Relatively regional similarity criteria could be made by partly restraining the feature space (i.e., clustering) into multiple regions; then, adaption samples can be chosen for these "regional" clusters, which are analyzed in the lab using also conventional measurements. Finally, data-driven selection is optimized for the subsets made using RS-LOCAL. This could lead to efficiency improvement in selecting relevant samples (quantity of samples and their characteristics), reduce computational needs, and improve the efficiency (less reference measurements and higher performance of spectroscopic estimates of soil characteristics). We will carefully revise this section again, to make concise statements about potential new directions/hypotheses.

RC2: P25L14. I've understood 209 variables were considered (cf. P6L29, and P25L18!)
AC Thank you. We will rectify the mistake.

RC2: P25L19. What are these model usage statistics?
AC These are for assessing the importance of each spectral variable, based on the usage of each individual variable in the rule conditions and the model for Cubist. We have described it in the methods, section 2.5.2. We will simplify "equally weighted usage of particular variable in split conditions and regressions " as outlined in our previous response (page 8, line 6–7).

4 References

Bui, E. N., Henderson, B. L., & Viergever, K. (2006). Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling, 191(3), 431–446. https://doi.org/10.1016/j.ecolmodel.2005.05.021

Dangal, S. R. S., Sanderman, J., Wills, S., & Ramirez-Lopez, L. (2019). Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library. Soil Systems, 3(1), 11. https://doi.org/10.3390/soilsystems3010011

Gubler, A., Wächter, D., Schwab, P., Müller, M., & Keller, A. (2019). Twenty-five years of observations of soil organic carbon in Swiss croplands showing stability overall but with some divergent trends. Environmental Monitoring and Assessment, 191(5), 277. https://doi.org/10.1007/s10661-019-7435-y

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., Zornoza, R., Ruiz-Sinoga, J. D., & Kuang, B. (2014). Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset: Spiking and extra-weighting to improve soil organic carbon predictions with NIR. European Journal of Soil Science, 65(2), 248–263. https://doi.org/10.1111/ejss.12129

Helfenstein, A., Baumann, P., Viscarra Rossel, R., Gubler, A., Oechslin, S., & Six, J. (2021). Predicting soil carbon by efficiently using variation in a mid-IR soil spectral library. SOIL Discussions, 1–29. https://doi.org/10.5194/soil-2020-93

Miller, B. A., Koszinski, S., Wehrhan, M., & Sommer, M. (2015). Comparison of spatial association approaches for landscape mapping of soil organic carbon stocks. SOIL, 1(1), 217–233. https://doi.org/10.5194/soil-1-217-2015

Pascucci, S., Casa, R., Belviso, C., Palombo, A., Pignatti, S., & Castaldi, F. (2014). Estimation of soil organic carbon from airborne hyperspectral thermal infrared data: A case study. European Journal of Soil Science, 65(6), 865–875. https://doi.org/10.1111/ejss.122

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning to predict soil properties from regional spectral data. Geoderma Regional, 16, e00198. https://doi.org/10.1016/j.geodrs.2018.e00198

Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., & Greve, M. H. (2015). Modeling Soil Organic Carbon at Regional Scale by Combining Multi-Spectral Images with Laboratory Spectra. PLOS ONE, 10(11), e0142295. https://doi.org/10.1371/journal.

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

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 131(1–2), 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

Viscarra Rossel, R. A., Brus, D. J., Lobsey, C., Shi, Z., & McLachlan, G. (2016). Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. Geoderma, 265, 152–163. https://doi.org/10.1016/j.geoderma.2015.11.016