Review 1

GENERAL COMMENT

The paper presents MIRS predictions of soil total C and N concentrations (TC, TN) in six regions of Central Africa separately, using the AfSIS Sub-Saharan library with no Central African soils (Strategy 1), possibly completed with the samples from the five other regions (Strategy 2), possibly completed with spiking samples from the same region (Strategy 3). This is done with the Memory-based learning (MBL) regression procedure, which uses spectral calibration neighbors for building a PLS regression for each target sample individually.

This is very interesting, but the paper suffers several drawbacks. Some methodological aspects are not presented (selection of the number of latent variables in global calibrations developed for optimizing spectral pretreatment; window size for calculating spectral similarity; possible cut-off value for spectral similarity; minimum and maximum number of latent variables for calculating weighted average predictions) or not discussed (pretreatment selection on X residues instead of Y residues, as usually; forcing spiking samples into neighborhoods; why not testing a strategy without AfSIS dataset, to evaluate its usefulness), some terms are not introduced/defined (hold-out and validation sets; MEpred; notion of accurate prediction), and some points are unclear (what were Central African samples out of the six core regions used for? why were AfSIS sentinel sites divided into hold-out and validation sets?). Some results are misinterpreted (using RMSE for comparing predictions between regions with different distributions of TC or TN; differences between strategies), others are not presented in the text (effect of the number of spiking samples) or not discussed (negative effects of Strategy 2 in several cases), and conclusions often seem too optimistic ("accurate predictions" etc. while error represented >=30% of observed mean in most cases).

For these reasons, I recommend moderate revision.

We thank the reviewer for the detailed comments and constructive criticism. We fully agree that there is some methodological information missing, which we will add accordingly or discuss in the corresponding comments below. We also acknowledge that some terms need further explanation and the structure of the presentation needs to be more clear. We will therefore only present the six regions we actually worked with in our manuscript and will add a table for the entire spectral library in the annex. We also agree that there were some misinterpretations that arose from considering only $RMSE_{pred}$. To analyse the predictions between the regions, we will use the $RPIQ_{pred}$ instead, and only when comparing strategies within the same region, we use the $RMSE_{pred}$. Additionally, we propose to replace Table 3 by a figure (see below), which visually depicts each distribution of total carbon and total nitrogen contents, including a boxplot, showing the interquartile ranges. Furthermore, we discuss the partially negative effect of strategy 2 compared to 1 in detail and also report and discuss the effect of spiking in detail. Overall, we are convinced that we will be able to implement these changes as discussed below and it will improve the readability and quality of the manuscript.

Please note that in the answers we provide here we do not distinguish between the comments made by the reviewer in capital letters and comments made in lowercase letters.

SPECIFIC COMMENTS

The title is short, which may be an advantage, but I wonder if it is informative enough; moreover the genericity of the work is not highlighted (i.e. using a large spectral library for predictions in poorly documented areas).

The actual title is a general title meant for a broad audience interested in quantitative soil assessments, who may not necessarily be experts in spectroscopy. However, we agree that the title should be more specific and will change it accordingly:

A new soil infrared library for central Africa and a geographical prediction analysis.

L8-9. What was done with the six core regions, and what the three levels of extrapolation consisted of, should probably be specified a little bit. Moreover, specifying the size of AfSIS SSL would be useful.

We agree with the reviewer that the strategies in the abstract lacked some details described in the abstract and will implement the following changes starting from line 7 (see below):

"*For the analysis, we used six regions from the CSSL, which we predicted using an existing continental SSL (African Soil Information Service, AfSIS; n = 1902) that does not include central African soils. We explored three different strategies, at decreasing degree of geographic extrapolation, to predict total carbon (TC) and total nitrogen (TN) contents of the six selected regions using models built with (1) the AfSIS SSL only, (2) AfSIS SSL combined with the five remaining regions, and (3) a combination of AfSIS SSL, the remaining five regions, and selected samples from the target region.*"

L13-14. Improvement was not clear for TC, from RMSE=0.38-0.86% to 0.41-0.89%. more details? split up into regions? Moreover, I wonder if such prediction errors allow considering the approach as particularly useful (i.e. is information ACCURATE ENOUGH?). Note that RMSE is not particularly informative as long as distribution has not been specified (e.g. RMSE=3 is small if mean=30 and SD=10, but high if mean=10 and SD=5), so adding RPIQ would be useful.

We agree with the reviewer that the presentation of these ranges does not show improvements. This is now clarified in the abstract (see above); we will change the abstract as proposed above and will suggest that readers assess the cost-benefit of investing in new sampling versus gaining accuracy.

L38. Cost is one reason, there are probably others.

The reviewer is correct, there are numerous other reasons for missing soil data in central Africa, including but not limited to accessibility to sampling areas, infrastructure, and political instability. We will include these other factors in the revision.

L52-53. The notion of "positive predictive transfer" is unclear for me.

We thank the reviewer for the comment and we agree that this notion was not clearly formulated. With “positive predictive transfer” we describe the information transferred from a large infrared library for a new calibration of a local set as described by Padarian et al. (2019). The calibration of a new local set using a large-scale spectral library can be complex in soil science, especially when the local set covers a different geographical domain than the library. Soil spectral libraries become particularly useful when a large amount of their relevant information can be extracted in a way that it improves prediction accuracy (positive transfer) and minimizes the number of additional costly local reference measurements for quantifying soil properties in the local set (accuracy-cost trade-off). To avoid technical jargon we will rephrase the paragraph L48-L55 and move it to L60, where it fits better into the context:

“One of the main aims of establishing large-scale SSLs is to minimize the need for future wet chemical analyses (e.g., Nocita et al., 2014; Stevens et al., 2013; Shi et al., 2014; Viscarra Rossel et al., 2016). However, these libraries often span vast geographical areas that include different soil types and climate zones, which comprise complex soil organic C forms and mineral compositions. Due to this heterogeneity, predictions rendered by global linear regression models are often unfeasible for new local soil property assessments at a regional, field or plot-scale, especially when the new set covers another geographical domain than the library. Pandiran et al. (2019) could considerably improve prediction accuracies for a new local set when using a compositionally related subset from a large-scale SSL together with a small number of local reference analyses. The cost-accuracy trade-off can be met when the accuracy of the library-based prediction is similar to the one made when applying a local but more costly calibration strategy. Several data-driven methods have proven to be successful to overcome this issue, for example RS-LOCAL (Lobsey et al., 2017) and memory-based learning (a.k.a local learning e.g. Ramirez-Lopez et al., 2013; Shenk et al., 1997; Naes 1990). In addition, other promising approaches have also been proposed, although they require more research (e.g. deep learning (Ng et al. 2019), fuzzy rule-based systems (Tsakiridis et al. 2019)).”

Padarian, J., Minasny, B., McBratney, A.B.: Transfer learning to localise a continental soil vis-NIR calibration model, Geoderma, 340, 279-288, https://doi.org/10.1016/j.geoderma.2019.01.009, 2019.

L64-67. LOCAL and Locally weighted PLSR should probably be cited, as they also aim at selecting spectral calibration neighbors, and were used earlier in soil spectroscopy.

We agree with the reviewer and propose the following changes, together with the next comment (L64-70) (see below).

L64-70. In my opinion, approach complexity should be considered: some approaches are rather simple (e.g. spiking) thus widely usable, while others are complex thus usable only by experts (e.g. the fuzzy rule-based system proposed by Tsakiridis et al. 2019).

We will rephrase the paragraph as following and add two references to the reference list:

“Several data-driven methods have proven to be successful in overcoming this issue, for example RS-LOCAL (Lobsey et al., 2017) and memory-based learning (a.k.a local learning e.g. Ramirez-Lopez et al., 2013; Shenk et al., 1997; Naes 1990). In addition, other promising approaches have also been proposed, although they require more research (e.g. deep learning (Ng et al. 2019), fuzzy rule-based systems (Tsakiridis et al. 2019)).”

Naes, T., Isaksson, T., & Kowalski, B.: Locally weighted regression and scatter correction for near-infrared reflectance data. Analytical Chemistry, *62*, 664–673, https://doi.org/10.1021/ac00206a003, 1990.

Tsakiridis, N., Theocharis, J., Panagos, P., & Zalidis, G.: An evolutionary fuzzy rule-based system applied to the prediction of soil organic carbon from soil spectral libraries. Applied Soft Computing, 81, 1-18, https://doi.org/10.1016/j.asoc.2019.105504, 2019.

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

L86. "covers a large geographic area" is questionable as the sample population is clustered, and a wide area is not represented (i.e. between Kinshasa, Tshopo and Katanga).

The reviewer is correct! The sampling locations did not cover the entire area and the term is potentially misleading. We will address this comment in line 86 accordingly:

“The sample locations are clustered in eight regions distributed over a large geographical area of central Africa, from a latitude of …”

L99. The way samples were dried should be specified, moreover they had probably been 2-mm sieved previously.

We thank the reviewer for requesting this information. The samples were all sieved through a 2 mm mesh and either air dried or oven-dried at temperatures of 50 °C, 60 °C or 105 °C, all of them suitable for total carbon and nitrogen analyses. After sieving and drying, soil samples were ground to a powder (< 50 µm) using a ball mill. We will include these details in the revised manuscript

Tab.2. I've not understood how samples from Equateur, Bas-Uélé, North Kivu and Kongo-Central were used (they are not mentioned in Strategy 2, L204-205).

The regions Equateur, Bas-Uélé, North Kivu and Kongo-Central were excluded for the further analyses because they did not have enough samples to allow for reliable analysis (< 80 samples per region). With this table, we intended to present the entire infrared library we created. However, we fully understand that this is confusing here and we will remove these regions from this table but present the full library (including these four regions: Équateur, Bas-Uélé, North Kivu and Kongo-Central) in a supplementary table in the appendix.

L106-107. Does this suggest charcoals were considered organic, or negligible?

This is a legitimate question, since slash-and-burn is commonly used to clear fields in central Africa which adds charcoal to the topsoils. For our soil analyses, visible pieces of charcoal were removed, which could clearly influence TC measurements in certain samples. This detail will be added in the methods.

L112. SPECIFYING PARTICLE SIZE WOULD BE USEFUL (< 0.2 mm? < 0.1 mm?).

All samples were grinded to a powder (<50 µm) using a ball mill, which is sufficiently accurate for soil spectral diagnostics. Diess et al. (2020) report sufficiently accurate model estimates when grinding below 0.5mm, and Guillou et al. (2015) even report no significant differences at particle size thresholds of 1.0mm, 0.5mm and 0.25mm thresholds. We will add this information to the method section.

Deiss, L., Culman, S. W., & Demyan, M. S.: Grinding and spectra replication often improves mid-DRIFTS predictions of soil properties, Soil Science Society of America Journal, 84, 914–929. https://doi.org/10.1002/saj2.20021, 2020.

Guillou, F. L., Wetterlind, W., Viscarra Rossel, R. A., Hicks, W., Grundy, M., & Tuomi, S.: How does grinding affect the mid-infrared spectra of soil and their multivariate calibrations to texture and organic carbon? Soil Research, 53, 913-921, https://doi.org/10.1071/SR15019, 2015.

L113, L125. Spectral range and resolution should probably be specified.

We thank the reviewer for bringing this to our attention. We fully agree and will change the sentences accordingly:

All samples were measured with a VERTEX70 Fourier Transform-IR (FT-IR) spectrometer with a High Throughput Screening Extension (HTS-XT) (Bruker Optics GmbH, Germany) in order to measure their MIR reflectance spectra. Spectra were acquired in a resolution of 2 cm^{-1} *within a range of 7500 cm*$^{-1}$ *to 600 cm*$^{-1}$*, which corresponds to a wavelength range of 1333 nm to 16667 nm. A gold coated reflectance standard (Infragold NIR-MIR Reflectance Coating, Labsphere) was used as a background material for all measured soils in order to normalize the sample spectra. Reflectance was transformed into absorbance using log(1/reflectance) prior to further processing and subsequent modeling.*

L125. Spectra were collected on AfSIS and CSSL samples with different spectrometers, so the question of compatibility should be addressed (e.g. was there standardization?).

The reviewer raises an important point regarding the compatibility of data form two different spectral libraries. Luckily, the two instruments were both FT-IR spectrometers from BRUKER which use the same settings and the same internal standards. The scanning methods of the CSSL were adapted to the ICRAF standard operating procedures. For these reasons, no instrument standardization was necessary and all spectra between the libraries can be compared one to one. This information will be added to the methods section of the revised manuscript.

L132. A reference dealing specifically with soils would probably be more appropriate.

We agree that a more soil specific reference would help to point out the importance of the effect of pre-processing and we therefore suggest the two following publications:

Seybold, C.A., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Nester, B., Schoeneberger, P., Wills, S., Libohova, Z., Hoover, D. and Thomas, P.: Application of Mid-Infrared Spectroscopy in Soil Survey. Soil Science Society of America Journal, 83, 1746-1759, https://doi.org/10.2136/sssaj2019.06.0205, 2019.

Sila, A. M., Shepherd, K. D., and Pokhariyal, G. P.: Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties, Chemometrics and Intelligent Laboratory Systems, 153, 92–105, https://doi.org/10.1016/j.chemolab.2016.02.013, 2016.

L140. p is not defined. Actually P is a d x l matrix, not a d x p matrix.

We thank the reviewer for spotting this typo! “d × l matrix” is correct and we will change it as suggested by the reviewer.

L145-161. The error E depends on the NUMBER OF LATENT VARIABLES (l). HOW WAS THIS PARAMETER DEFINED? Moreover, the EXPECTED BENEFIT OF THIS APPROACH (i.e. computing Xcssl residues) for optimizing spectral pretreatment SHOULD BE PRESENTED, when compared with examining RMSE associated with every pretreatment (i.e. computing Ycssl residues, as commonly done).

We fully acknowledge that this was not clearly explained in the text and will address these issues. We explain that the analysis of spectral reconstruction error is indeed commonly used in spectroscopy for outlier identification. This error is also known as the Q-statistic and it indicates how well a given new sample conforms to the PLS model. Since the response values in the prediction set are unknown, we can use the Q-statistic as a proxy for the response errors. In the revised version, we will explain that we assume that if a given set of pre-processing steps lead to large Q-values, then it is expected that it will also lead to large errors in the prediction of the response values. We will also add references to support this assumption. In the new version of the manuscript, we will mention that for this analysis we fixed the number of PLS factors to 20, as projected variables beyond this dimension did not capture a considerable amount of the original spectral variance. For example, PLS variable 21 amounted for less than 0.01% of the original variance in all the cases.

L165."spectral matrices which can be properly represented by a PLS model" is unclear. Moreover, the assumption that SIMILAR PRETREATMENTS OPTIMIZED GLOBAL AND LOCAL CALIBRATION SHOULD BE DISCUSSED (e.g. according to literature).

The ideas behind this sentence will be clarified with the description of the Q-statistic (see previous reply L145-161) and the advantages of its use for pre-processing optimization. We indicate now that according to Wise and Roginsky (2015), large Qc values are proxies to large prediction errors and therefore Q-statistic can be used to judge the suitability of a set of pre-processing steps.

Wise, B. M., & Roginski, R. T.: A calibration model maintenance roadmap. IFAC-PapersOnLine, 48, 260-265, https://doi.org/10.1016/j.ifacol.2015.08.191, 2015.

L170. The problem with multiplicative scatter correction is that the transformed spectrum depends on the spectrum population it belongs to, so changes when this population changes.

We thank the reviewer for raising this concern but do not see this as a problem. Multiplicative scatter correction (MSC) aligns or rotates a given spectrum towards a reference one which is fixed. This reference spectrum can be seen as a parameter of the MSC transformation. By doing this, multiplicative and additive shifts between spectra are removed. Although, in many applications the average spectrum of the calibration set is used as the reference one, in theory any spectrum can be used (See Rinnan et al., 2009). Therefore, MSC is not necessarily affected by changes in the spectral population. The reference spectrum parameter of a defined MSC step should not be modified as long as it guarantees successful removal of the multiplicative and additive scattering effects across the spectra.

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

L195. Why 20 spiking samples per regional set, not 10 or 30?

We agree with the reviewer, that the selection of the number of spiking samples has not been adequately described in the manuscript. Generally, the number of spiking samples should be minimized to reduce costs for laboratory reference analyses. We set the maximum number of spiking samples to 20, which can already mean quite a high financial investment but we feel that it is worth these costs given the reduction of geographical extrapolation and the effect of using spatially close samples on the predictive performance. We tested one to 20 spiking samples and compared the prediction accuracy, which was on average best with 20 spiking samples (Figure 5). We will add more details about the spiking effect in the results and discussion sections of the manuscript.

L197. The way k-means works could (should?) be briefly presented.

Since this method is widely used and well documented in pedometrics and chemometrics for sampling calibration datasets, we considered that it was sufficient to refer the reader to other studies, where k-means sampling is explained. However, we agree that it is useful when we explain it in a sentence and will change it as following:

"... for each complete regional set, 20 samples were selected using the k-means sampling algorithm. This sampling strategy is implemented in the R package prospectr (Stevens and Ramirez-Lopez, 2020) and selects one sample per cluster calculated with a k-means algorithm on a principal component analysis of the pre-processed spectra (Næs, 1987).

L199. The strategies considered are: AfSIS alone; AfSIS +other Gi; AfSIS +other Gi +Ki. Other strategies would have been interesting: only using other Gi, or other Gi + Ki, to EVALUATE THE USEFULNESS OF AfSIS (which would be very interesting); AfSIS +Ki, to evaluate the usefulness of other Gi; Ki only, to evaluate the usefulness of AfSIS and other Gi. But this would require much additional work!

Our aim was to propose strategies that could leverage the use of the AfSIS spectral library to accurately predict soil properties in regions which are poorly covered by it. Therefore, we only evaluated modeling approaches that involved the use of this library. There is clear evidence that very accurate soil predictions can be achieved by using models built only with samples originating from the same region or area where these predictions are required. This is because large non-linear complexity is avoided in local-scale models (See e.g., Tziolas et al., 2019). Despite this, we consider that this implies that every undersampled region will require a representative calibration sample set which might be expensive or impractical. In this respect, the evaluation of models using only other Gi, or only other Gi + Ki was not considered as they do not really solve the problem of using a large spectral library in poorly sampled areas.

Tziolas, N., Tsakiridis, N., Ben-Dor, E., Theocharis, J., & Zalidis, G.: A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation,Geoderma, 340, 11-24, https://doi.org/10.1016/j.geoderma.2018.12.044, 2019.

L217-219. HOW WAS w DEFINED? Moreover, WHAT p STANDS FOR IS NOT CLEAR: it has not been defined, but according to L140, was apparently used in place of l (number of latent variables); but I'm not sure this makes sense here. Furthermore, I'm not sure to understand what k=1 means. I also note that d has already been used (number of wavelengths; L139). So CLARIFICATION IS REQUIRED. We might also wonder why evaluate dissimilarity (1-S) and not similarity (S), when the objective is to select calibration samples *similar* to the target sample (cf. L311). Furthermore, I WONDER IF A SIMILARITY/DISSIMILARITY CUT-OFF VALUE WAS DEFINED, below/above which spectra were not considered neighbors (i.e. no prediction for target samples with too few neighbors); and if yes, how this cut-off value was defined.

We are thankful that the reviewer noticed the use of letters for multiple variables, which is misleading. Again, the reviewer is correct, spotting the mistake in L140, which leads to confusion in L215-219. Correcting this as suggested above, this issue should be resolved here.

The window size (w) was optimized based on a spectral nearest-neighbor search within the AfSIS library. For every sample in the AfSIS library, its closest sample (in the spectral space) was identified. The samples were compared against their closest samples in terms of TC and TN and the root mean squared differences (RMSD) were computed according to the following equations:

$$j(i) \; = \; NN(Xc_i, Xc^{-i})$$

and

$$RMSD \; = \sqrt{\frac{1}{2m}\sum_{i=1}^{m} \; \sum_{h=1}^{2} \; (yc_{i,h} - yc_{j(i),h})^2}$$

where $Xc$is the spectra of the AfSIS library, $NN(Xc_i, Xc^{-i})$ represents a function to obtain the index of the nearest neighbor observation of the ith sample found in Xc(excluding the ith sample), $yc_{i,h}$is the value of the i-*th* observation for the *h*-th property variable (either TC or TN). A total of 10 window sizes were evaluated (from 31 up to 121 in steps of 10). According to the RMSDs obtained, the optimal *w* was 71.

Concerning the comment about using the concept of similarity or dissimilarity, we believe that is not actually relevant. It is clear that similarity or dissimilarity measures can be both used to identify similar samples. Many examples of the use of correlation dissimilarity for nearest neighbor identification can be found in the NIR spectroscopy literature (See for example Wadoux et al., 2021; Khosravi et al., 2020; Gholizadeh et al., 2018; Zhu et al., 2011).

Gholizadeh, A., Saberioon, M., Carmon, N., Boruvka, L., & Ben-Dor, E.: Examining the performance of PARACUDA-II data-mining engine versus selected techniques to model soil carbon from reflectance spectra. Remote Sensing, 10, 1172, https://doi.org/10.3390/rs10081172, 2018.

Khosravi, V., Ardejani, F. D., Aryafar, A., Yousefi, S., & Karami, S.: Prediction of copper content in waste dump of Sarcheshmeh copper mine using visible and near-infrared reflectance spectroscopy. Environmental Earth Sciences, 79, 1-13, https://doi.org/10.1007/s12665-020-8901-0, 2020.

Wadoux, A., Malone, B., Minasny, B., Fajardo, M., McBratney, A.B. (Eds.): Soil Spectral Inference with R: Analysing Digital Soil Spectra using the R Programming Environment, Springer Nature, Cham, Switzerland, 2021.

Zhu, Z., Corona, F., Lendasse, A., Baratti, R., & Romagnoli, J. A.: Local linear regression for soft-sensor design with application to an industrial deethanizer, IFAC Proceedings Volumes, 44, 2839-2844, https://doi.org/10.3182/20110828-6-IT-1002.02357, 2011.

L220-225. According to Shenk et al. (1997), the weighted average is calculated over a range of latent variables, i.e. from a MINIMUM TO A MAXIMUM NUMBER OF LATENT VARIABLES CONSIDERED, AND THESE PARAMETERS HAVE TO BE SPECIFIED. Moreover, both s1:j and gj are calculated for the jth latent variable, so writing "s1:j" instead of "sj" is unclerar. Furthermore, Shenk et al. (1997) did not call this approach "Weighted averaged PLS"; but why not…

As correctly pointed out by the reviewer, details about the WA-PLS are missing in the current version of the manuscript. We will add the missing information to the text. The weighted average was calculated using a range of latent variables from 5 to 30 in increments of 1, which we will add to the manuscript accordingly.

To compute the weights we use the exact same method as described by Shenk et al. (1997, see page 227 of their paper). In the equation used to compute the weights, s1:j represent the root mean square of the spectral residuals of the query spectrum. The reconstruction is done by multiplying the scores of the projected query spectrum by the (transposed) loading matrix of the PLS model built from its neighbor samples. In this multiplication the first *j* rows of the scores and loading matrices are used. Using sj instead of s1:j, would wrongly indicate that only the *j*th row of the scores is multiplied by the *j*th transposed row of the loadings. Furthermore, in the equation we also use the term gj to refer to the root mean square of the regression coefficients corresponding to the *j*th PLS component. In this case we do not use the subscript 1:j as we are using only the *j*th row of the matrix of regression coefficients (instead of the first *j* rows). We will extend the explanation of this notation for a new version of the manuscript.

Indeed Shenk et al., (1997) do not explicitly call this regression method “weighted averaged PLS”. Although, what this method does is to compute a “weighted average of the individual model predicted values with from the minimum to the maximum number of factors” as explained by Shenk and Westerhaus (1998) in the following patent filing: https://patents.google.com/patent/US5798526A/en. Therefore, we do not see the term “weighted averaged PLS” as incorrect in our manuscript.

L230-232. Hold-out and validation sets have not been introduced, so this part is not very clear (e.g. why dividing regional AfSIS sub-libraries into hold-out and validation sets? L256 and Tab.3 these sub-libraries were not separated).

We thank the reviewer for spotting this point of confusion. We will clarify this issue as following:

The grouping factor was used for the optimization of the nearest neighbor search, i.e. the nearest neighbor cross-validation (see L226) to *avoid overfitting: keeping the nearest neighbor out, the model was trained with the remaining neighbors which were not from the same region as the hold-out neighbor (region corresponds to the sentinel sites within the AfSIS SSL).*

This will be changed accordingly to avoid a misunderstanding as shown in this comment.

L233. I understand the minimum requested number of neighbors was 150, and the maximum possible number of neighbors was 500. WHAT IF A TARGET SAMPLE HAD LESS THAN 150 NEIGHBORS?

This is an important question of the reviewer. Of course, a sample could have less than 150 neighbors in the used spectral library. We tested the minimum number of available neighbors prior training the final model. We agree that the minimum number of neighbors should have been adjusted downwards if there would not have been enough neighbors, which was luckily not the case. We will explain this more in detail in the manuscript.

L236. FORCING SPIKING SAMPLES INTO THE NEIGHBORHOOD of every target sample is questionable, and the discussion should address this point.

Unfortunately the reviewer does not provide an explanation on why forcing spinking samples into the neighborhood is questionable.

Spectral Neighbor identification is a mathematical attempt to select soil observations that share similar compositional characteristics with the observation that requires a prediction. MIR spectra partially reflect the compositional characteristics of the samples. We assume that soils originating from the same geographical region might be governed by very similar soil formation processes. This is a concept of spatial autocorrelation which is widely used (Fortin et al. 2016). Furthermore, it is widely accepted that the best spectral models (most accurate) that can be built for a given area are those that are calibrated with samples from the same area (see also comment L199). For these reasons, we assume that forcing samples of a given area to belong to the neighborhoods of samples from the same area guarantees that samples originating from similar soil formation processes are included in the models. Therefore, our approach is not arbitrary as it is expected that these samples improve prediction accuracy.

Fortin, M.-J., Dale, M.R. and Ver Hoef, J.M.: Spatial Analysis in Ecology. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat07766.pub2, 2016.

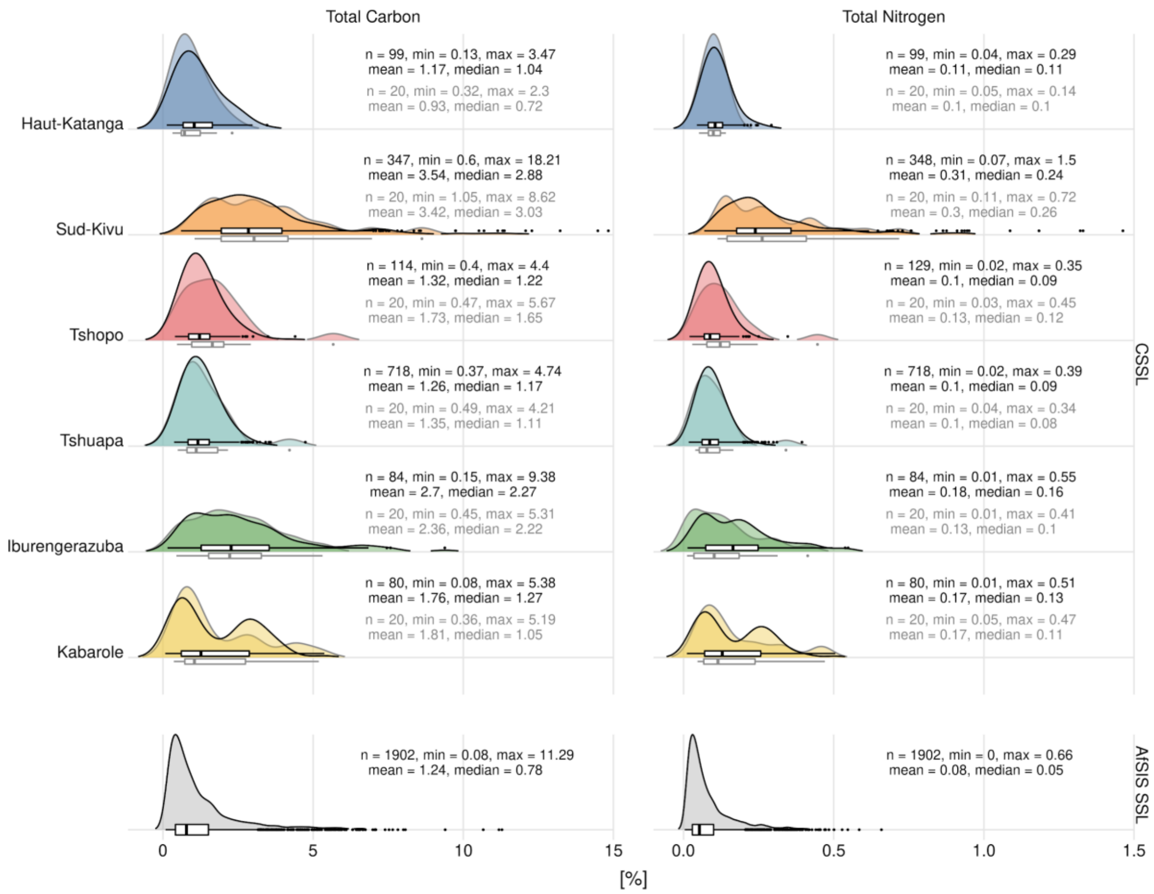
Fig.3. Beside orange and green circles, many grey circles were also outside AfSIS black circles, and it would be useful to mention where they originated from.

The transparency of the black AfSIS symbols is misleading. They seem gray, while the remaining samples are black, where the density is high. We will increase the transparency and change the style of the symbols, so that it becomes clear which points belong to the AfSIS library.

L265-267. CRITERIA FOR "GOOD PREDICTIVE RESULTS" HAVE NOT BEEN SPECIFIED. Actually many results were not so good, especially for TN, especially with Strategy 1 (e.g. RMSE for TC and TN was >=50% of observed mean for 2-3 regions with Strategy 1, and >=30% of the mean for 4-5 regions with Strategy 2). And ACCORDING TO RPIQ, PREDICTIONS FOR SOUTH KIVU AND IBURENGERAZUBA WERE OFTEN AMONG THE BEST ONES, so the reasons for considering they "showed the lowest accuracy levels" should be revised, or at least explained.

The reviewer is correct, we have not introduced criteria to define a good or accurate prediction which we will add. However, the required prediction accuracy depends on the field of application. We will work on a method to assess prediction accuracy for a hypothetical new sample set.

As the reviewer points out, $RMSE_{pred}$ is useful to estimate prediction accuracy within the same region but not to make comparisons between regions since ranges for TC and TN differ between the regions, especially for Iburengerazuba and South Kivu with forest soils with high TC and TN contents. We agree that it does not make sense to classify the statistical performance of the South Kivu and Iburengerazuba regions as poor. Indeed, when looking at the $RPIQ_{pred}$ values, they performed well. This is due to the large interquartile (IQ) range of these regions compared to Tsuapa and Tshopo, which exhibited considerably smaller IQ ranges. We thank the reviewer for this careful attention to these statistical descriptions and will modify the results and discussion accordingly. Moreover, we will replace Table 3 with the proposed plot (see below), which clearly shows the distribution of TC and TN in each region including their IQ ranges in the boxplots. They cray coloured line and text indicate the spiking sets (K_i), the black coloured lines and text represent the six regional sets (*Gi*) after removal of each K_i and the AfSIS SSL data set *A*. These details to the figure will be added to the caption.



L271-272. RMSEpred is useful for comparing strategies for a given region, but CANNOT BE THE FIRST PARAMETER CONSIDERED FOR COMPARING PREDICTION ACCURACY BETWEEN

REGIONS WHERE DISTRIBUTIONS OF TC OR TN WERE DIFFERENT. R^2 describes proportionality, not similarity; so, though understood by a wide audience, should be used with care. Comparison between regions should firstly be based on RPIQ, which showed good results for Kabarole, Iburengerazuba and (for TC) South Kivu and poor results for the other regions, especially Tshopo for TC and Haut-Katanga for TN.

Yes we agree and as we detailed in the response above, we will modify the results and discussion such that we only use $RMSE_{pred}$ to compare the same regions across strategies and $RPIQ_{pred}$ and R^2_{pred} to compare regions within a given strategy. We suggest the following changes:

"The best prediction accuracies for TC were achieved for the regions South Kivu, Iburengerazuba and Kabarole, where $RPIQ_{pred}$ values were between 2.43–3.95, while Tshopo, Tshuapa and Haut-Katanga performed less good with $RPIQ_{pred}$ <= 1.84. For TN, Iburengerazuba and Kabarole performed well with $RPIQ_{pred}$ 2.14 and 2.86, respectively. However, the four other regions Haut-Katanga, South Kivu, Tshopo and Tshuapa exposed smaller $RPIQ_{pred}$ <= 1.37. "

L277-279. The fact that CENTRAL AFRICAN SAMPLES WERE POORLY REPRESENTED BY AfSIS SHOULD ALSO BE MENTIONED AS POSSIBLE REASON.

We agree with the reviewer that this should be highlighted at this point and we will add this accordingly. We also suggest to put Figure A1 (continental map) in the main text and move Figure 1 to the appendix.

L282-283. Again, RMSEpred should not be used for comparisons between regions.

We agree with the reviewer and will change it as described more in detail in L271–272. We suggest the following changes:

"The predictive performance in strategy 2 exhibited errors ($RMSE_{pred}$) ranging between 0.41–0.89 % and 0.03–0.12 % for TC and TN, respectively (Table 4). The most accurate predictions for TC were as in strategy 1 obtained for the regions Iburengerazuba, Kabarole and South Kivu ($RPIQ_{pred}$ > 2.36), but $RPIQ_{pred}$ value of Haut-Katanga was remarkably higher than in strategy 1 (2.30 vs 1.62). Predictive performance for TC of Tshopo and Tshuapa were still below an $RPIQ_{pred}$ of 2.

For TN, similarly to strategy 1, prediction accuracy was good for Iburengerazuba and Kabarole. For the regions Haut-Katanga, South Kivu, Tshopo and Tshuapa the $RPIQ_{pred}$ values were higher than in strategy 1, but they were still below 2. "

L284-286. RMSEpred for TC increased in three regions from Strategy 1 to 2, strongly sometimes, which is counter-intuitive so should be underlined, and POSSIBLE REASONS SHOULD BE PROPOSED (as was done for better TN predictions with Strategy 2 than 1).

The reviewer is correct in that $RMSE_{pred}$ increased for 3 regions, however it only increased by 0.03% in two of the cases. So, in total, from Strategy 1 to 2 the $RMSE_{pred}$ decreased substantially in 3 regions, barely changed in 2, and increased in one, which in our opinion signals an overall improvement in performance. At the moment, it appears that the inclusion of the additional CSSL regions reduced the accuracy of the Kabarole region but it is unclear why the model did not fall back on the same prediction subset as Strategy 1. This will be investigated and corrected in the revised manuscript.

L287. Better TN predictions with strategy 2 than 1 "was due", not "might be due".

Thank you, this will be modified accordingly.

L290. RPIQ for TC "tended to be the same" except for Kabarole; but actually RPIQ decreased in South Kivu and Tshuapa, not much, but this is counter-intuitive.

As detailed above, we will modify the discussion of these results in the text to explain the observed patterns.

L292. South Kivu was not an exception, as TN prediction was also improved.

Thank you for this correction. We will modify the text accordingly.

Fig.5. THESE RESULTS SHOULD BE PRESENTED in the text, and an optimal number of spiking samples could be proposed for each region.

We thank the reviewer for requesting that these results be included in the text and will add them to the revised version.

L309, L317, L391. "Accurately predicted/model" "highly accurate predictions" are OVEROPTIMISTIC, e.g. when RPIQ <2 or RMSE > mean/2.

We thank the reviewer for this suggestion and will tone down the language to “reasonably accurate”.

L317-318. The point is that for TC, Strategy 2 reduced RMSEpred in only 3 out of the 6 regions considered; so "improved prediction accuracy" is questionable. And POOREST PREDICTION WITH STRATEGY 2 than 1 FOR 3 REGIONS SHOULD BE DISCUSSED.

We again thank the reviewer for pointing out this idiosyncrasy and as detailed in the responses above will modify the results and discussion to detail these prediction results.

L322-325. There is STRONG MISINTERPRETATION, as in these two regions, TC (and TN in Iburengerazuba) was accurately predicted (RPIQ >2.3).

As detailed above, these discussion points surrounding the prediction results will be modified.

L338. These results have not fully presented in the results section.

We thank the reviewer for pointing this out and will detail the spiking results in the results section.

L339. Three regions are cited, not two. Moreover, Strategy 3 yielded highest RPIQ whatever the region for both TC and TN; and the improvement was strong sometimes, with 10 spiking samples only (Kabarole and Iburengerazuba).

We thank the reviewer for pointing out this mistake. The text should read “three regions” and we will remove the word “somewhat” to reflect the strong improvement. We will further modify this section to say that Strategy 3 had a positive effect on all regions but an even stronger effect on the three regions we originally listed.

L343-344. For TN in South Kivu, RPIQ increased from 1.1 to 1.6 from Strategy 1 to Strategy 2, so prediction was noticeably improved.

We thank the reviewer for clarifying this point and will modify the text to say how the prediction noticeably improved.

L345. "RMSE remained relatively high", but TC and TN were much higher than elsewhere! Considering RMSE without considering TC and TN distributions leads to misinterpretation.

Indeed, the reviewer is correct. We will contextualize the $RMSE_{pred}$ with the higher TC and TN and instead focus on the $RPIQ_{pred}$ as a more reliable indicator given the different distributions. We also use the new graph to show this distribution of TC and TN for each region more precisely (see above).

L345. "slightly" does not seem appropriate: e.g. for Iburengerazuba RPIQ increased from 2.8 to 3.6 for TC and from 3.2 to 4.5 for TN.

We agree with the reviewer that “slightly” is not the correct word and will change it to “substantially”.

L349. As said above, the effect of spiking was strong sometimes (Iburengerazuba and Karabole).

We thank the reviewer for pointing this out and will modify the text accordingly.

TECHNICAL CORRECTIONS

L6. 1800 soils or 1800 soil samples?

Soil samples. We will clarify this in the text.

L7. "wider" is not clear for me in "Congo Basin and wider African Great Lakes region".

We will remove the word “wider” from this sentence. Moreover, we will correct “African Great Lakes region” to Albertine Rift, which is a more precise name for the region.

L10. % is not a SI unit and may cause confusion for comparisons or changes (e.g. TC increased by 5%), so G KG-1 WOULD BE MUCH PREFERABLE.

We thank the reviewer for this comment. We will convert the % unit into the SI unit g kg^{-1} as suggested by the reviewer.

L59. sol vs. soil.

Thank you for pointing out this typo.

L77. Predicting a region is confusing.

The reviewer is correct, this sentence does not make sense. We will specify accordingly in the updated version of the manuscript:

"... (2) To establish a workflow to accurately predict soils from variable locations within six selected geographical regions of the CSSL ...

L84. The sentence should be checked (e.g. layers vs. layer).

Thanks for spotting this typo, we will correct the word to the plural form.

Tab.1. Université catholique de Louvain and IITA/ICRAF are not references. Moreover, for the last reference, 2021a,b would be more appropriate than 2021b,a (this is detail).
We thank the reviewer for pointing out this detail. Indeed Université catholique de Louvain, IITA and ICRAF are not references. We will remove them and add an additional column named "Data Contributor".

L103. Total Al, Fe, Ca etc., or some particular fractions?

Total contents of cations have been analysed using aqua regia extractions. We agree with the reviewer that this should be specified and will add information about the methods for analysing pH, texture and cations.

L115. In general absorbance = log(1/reflectance), not 1/reflectance.

The reviewer is of course correct about this and it will be corrected accordingly.

L118. I note the manufacture place is mentioned here, which should probably be systematic.

This is correct, we thank the reviewer for seeing this detail. We will remove the place to be consistent through the entire manuscript.

L134. Actually PLS has most often been defined as Partial least squares.

The reviewer is correct, PLS is an abbreviation, used for Partial Least Squares. We also defined the term accordingly in the manuscript (L56). With the sentence the reviewer brings up, we do not want to give PLS another meaning. We rather want to explain that the Partial Least Squares method can also be described as a projection of latent structures, which has by accident the equal letters and the same order.

L207. The sentence should be checked.

We agree with the reviewer and will make the sentence clearer.

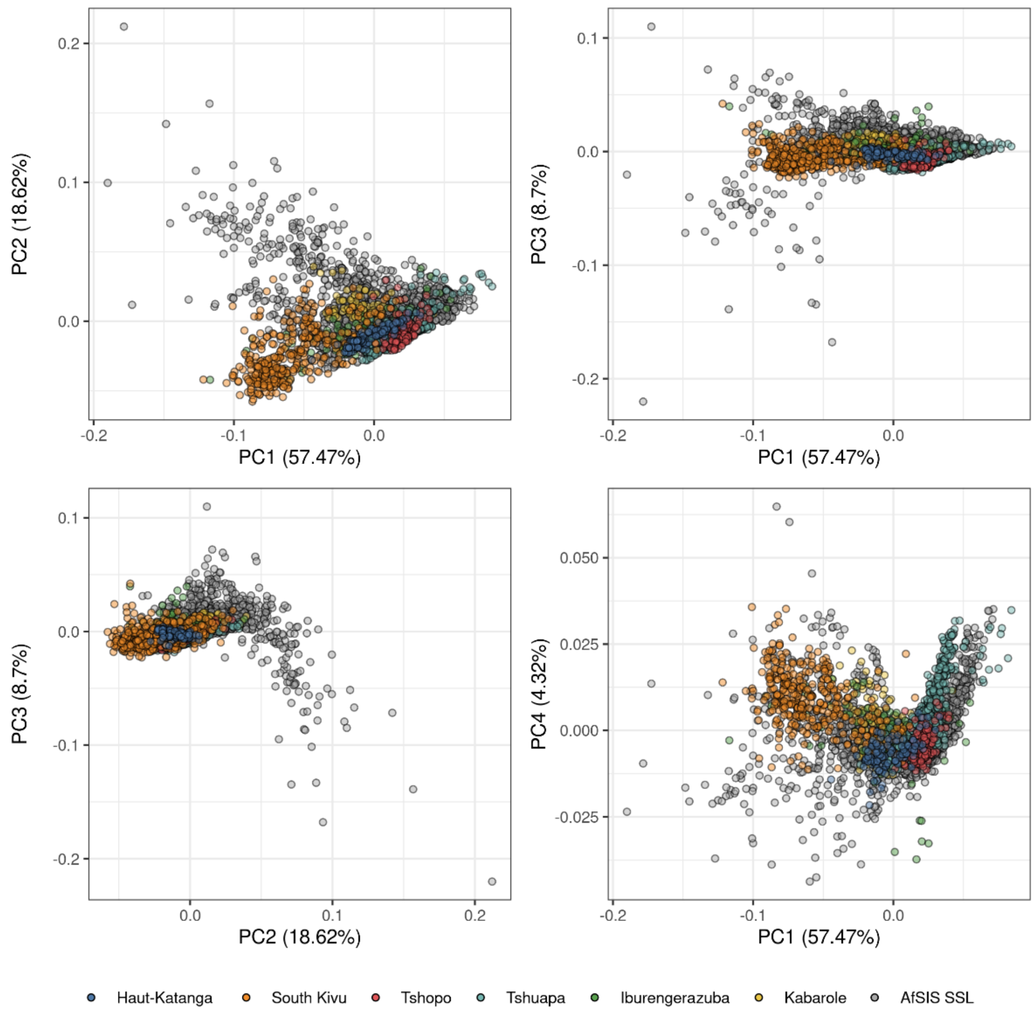
"– Strategy 3: This time, strategy 2 was repeated, but in this case, extrapolation was avoided by using the spiking samples from the same geographical region as the region to be predicted;"

L234, L243. Equation 8? Equations have not been numbered.

We thank the reviewer for spotting this mistake. We will remove "Equation 8" from the text.

Fig.3 is not very readable; projections on PC1-PC2 and PC1-PC3 would probably be more suitable.

We agree with the reviewer, that the 3D plot is not appropriate. We will therefore plot the three different components as suggested by the reviewer.



Tab.4. What MEpred stands for should be specified.

This is correct, we did not specify ME and will add this information to the methods section.

L275. Tshopo, not Tschopp. Four regions are cited, not three.

We thank the reviewer for clarifying this, and will change it accordingly.

L426-427, L432-433, L436, L439, L445, etc. Are two DOIs or two URLs necessary? I note that non-DOI URLs do not always work ("error 404", "page not found", etc.).

We thank the reviewer for checking the DOIs and URLs in the reference list. We will check them carefully.

L442, L445, L508, L540, L564, L567, L570, L573, L584-585, L615, L617-618. Same (or almost same) DOI mentioned twice.

We will also check these DOIs and remove the ones, which are not necessary. We thank the reviewer for spotting this issue.

L448, L469, L473, L485, L512, L599. DOI should be added.

Missing DOIs will be added to these references.

L482. What ISMEJ is should be specified.

We will specify this abbreviation, which is “Multidisciplinary Journal of Microbial Ecology”.

L487, L498, L530, L590, L591, L593, L611. The references do not seem complete.

We thank the reviewer for this comment and will add missing information to these references.

L501. European Commission Edn? Soil Atlas Series?

We thank the reviewer for spotting this typo and will correct the reference as requested in the corresponding document:

Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Micheli, E., Montanarella, L., Spaargaren, O., Thiombiano, L., Van Ranst, E., Yemefack, M. , Zougmoré R., (Eds.): Soil Atlas of Africa, European Commission, Publications Office of the European Union, Luxembourg. 176 pp, 2013.

L530. The publisher should be specified.

The publisher is Geoderma and we will add it accordingly, we thank the reviewer for spotting this issue:

Mujinya, B. B., Mees, F., Boeckx, P., Bodé, S., Baert, G., Erens, H., Delefortrie, S., Verdoodt, A., Ngongo, M., and Van Ranst, E.: The origin of carbonates in termite mounds of the Lubumbashi area, D.R. Congo, Geoderma, 165, 95-105, https://doi.org/10.1016/j.geoderma.2011.07.009, 2011.

613. This reference does not seem at the right place (Vagen et al. after Vollset et al.).

Following the Danish/Norwegian alphabet, “å” follows “z”. Therefore “Vågen et. al.” is at the correct alphabetic position after “Vollset et al.”.

L615. The end of the reference should be checked.

The reviewer is correct, the end of this reference includes some unnecessary information, which we will remove.

L622. I.W.G.?

I.W.G stands for IUSS Working Group WRB. We will correct this in the reference and replace it by the more recent version:

IUSS Working Group WRB: World Reference Base for Soil Resources 2014, update 2015 International soil classification system for naming soils and creating legends for soil maps, World Soil Resources Reports No. 106, Food and Agriculture Organization of the United Nations, Rome, Italy, pp. 193, ISBN978-92-5-108369-7, 2015.