REFEREE 1

In their manuscript the authors present the uncertainty of total carbon, soil inorganic carbon and soil organic carbon measurements depending on sample processing and measurement. The authors show substantial differences that are mainly driven by sieving and measuring methods with LOI being highly variable. It is or great importance to have such comparisons and critical assessments. The need for accurate soil C measurements is getting more important for an evolving C market. A substantial overestimation of C changes would be bad for the actual climate effect and a substantial underestimation of soil C would reduce the economic benefit of the C market. The experimental set up using 11 procedures and the comparison with 8 commercial laboratories is an important approach to reach a better homogenization of analyses approaches and make soil C measurements more consistent.

> **RESPONSE:** Thank you for your time and thorough comments. We appreciate your recognition of the merits of our study. Your feedback is very valuable and we hope that by addressing your comments, we have improved the manuscript for publication if the editor allows.

I have two main concerns:

The authors need to elaborate their discussion on the application of chemometric approaches by combining MIR and predictive modelling (e.g. Line 505-512, 523-526 and 567-570) It is true that such approaches can work well as reported in the cited literature. However, it needs to be clear that this all depends on the availability of a representative soil spectral library that it large enough to develop models for prediction. The good prediction in this study is expected and bias the generalized conclusion. The model was trained on the KSSL and thus covers the spectral variability of the soils used here. Additionally, the sample pre-treatment was very similar between the P0 method here and the initial data for the model presented in Seybold et al (2019). Seybold et al (2019) measured TC by dry combustion, used the pressure transducer method for SIC and determined SOC by difference. Thus, it is a good model for the soils selected here. However, the transferability of such models is difficult and a major challenge to overcome. For example, sample grinding is important for the transferability. Grinding was also an aspect that motivated the authors to test. It is true, the differences in grinding are not so significant when all samples are similarly prepared and the model trained for the corresponding grinding is applied but transferring a model trained on finely milled sample to samples that are coarser and vice versa brings uncertainty and challenges (Sandermann et al., 2023). Recently, Safanelli et al. (2023) reported that even combining spectra obtained from different devices can be difficult and requires important pre-processing. More importantly, the authors report that sample processing resulted in larger uncertainty of the predictions. Therefore, the authors need to constrain their conclusion here that such approaches are only working when the conditions of a good and regional model are given. Otherwise, the model error (e.g. RMSE) will be too large to detect changes in TC, SOC and SIC.

> **RESPONSE:** Thank you for turning our attention to some of the reasons why we obtained such accurate results from the FTIR analyses. We agree with these concerns and are sorry for having omitted discussing these limitations in the current manuscript. We will extensively and clearly present limitations in the revised manuscript, both in the methods and the discussion of results, as most appropriate. Specifically, we will present the advantage of having the KSSL spectral library available and reiterate the importance of having a robust library of samples derived from the same region of study, built on samples pre-processed with similar approaches, and scanned with the same protocol/instrument when using the FTIR approach to estimate soil properties. We will also refer to Safanelli et al (2023) to point to the difficulties associated with FTIR predictions when these conditions are not met.

We will revise the manuscript text as detailed below:

Current text (L23): The test suggested that the < 180 µm grind was sufficient for FTIR scanning and we used that for the comparison of P8 to the other quantification methods

Proposed text: The test suggested that the < 180 µm grind was sufficient for FTIR scanning, which was also the particle size of the samples used to build the NRCS-KSSL spectral library (Seybold et al., 2019) used in this study. Thus, we compared the Q1 < 180 µm protocol to the other quantification methods."

***P8** becomes Q1 in the revised manuscript as detailed below in response to the next comment

Current text (509): The level of fine grinding needed to obtain the most accurate and precise data from FTIR spectroscopy is unclear as results have been contradictory (Wijewardane et al., 2021; Sanderman et al., 2023). In our study, FTIR predictions were affected by the particle size after grinding for % SIC and % SOC, but not for % TC (Supplemental Fig. S1). The FTIR spectroscopy method may thus be a good alternative to EA as it is both reliable and more time and cost efficient.

Proposed text: The level of fine grinding needed to obtain the most accurate and precise data from FTIR spectroscopy is unclear as results have been contradictory (Wijewardane et al., 2021; Sanderman et al., 2023). However, Sanderman et al. (2023) showed that the level of grinding did not matter if the models were built from soils that were ground to the same particle size. This observation was confirmed by our work, as we observed that grinding to < 180 µm, which is the particle size of the NRCS-KSSL spectral library (Seybold et al., 2019) we used to build our FTIR models, was sufficient to obtain reliable predictions. In our study, FTIR predictions were affected by the particle size after grinding for % SIC and % SOC, but not for % TC (Supplemental Fig. S1). The FTIR spectroscopy method may thus be a good alternative to EA as it is both reliable and more time and cost efficient. It is worth noting that we obtained accurate results for the FTIR method because we used the same protocols and the same instrumentation for scanning our soils as was used to build the NRCS-KSSL library. Building models using samples processed differently or analysed using different instruments may have produced different results.

Current text (L525): These results indicate that calculating % SIC as % TC - % SOC (P9) is not as precise as quantifying % SIC directly using either predictions via FTIR or using a pressure transducer. Most importantly, it's crucial that testing for presence of SIC is incorporated into the standard operating procedures for soil processing in all soil testing labs, and that accurate quantification of SIC is carried out where its presence is detected. By not quantifying the inorganic C in calcareous soil, labs are overestimating true % SOC.

Proposed text: These results indicate that calculating % SIC as % TC - % SOC (Q2) is not as precise as quantifying % SIC directly using either predictions via FTIR or using a pressure transducer. As mentioned above, the accuracy of the FTIR method depends on the correspondence in terms of protocols and instrumentation between the samples analysed and those used to build the library (Safanelli et al. 2023). It is thus recommended that laboratories intending to use the FTIR method apply the same protocols used to build the library they intend to use for their prediction models.

***P9** becomes Q2 in the revised manuscript as detailed below in response to the next comment.

> **Current text (L567)**: We recommend the use of FTIR spectroscopy, particularly for SIC quantification as this method performed better than acid fumigation. Finally, we do not recommend LOI to measure % SOC and instead recommend the continued use of EA-PT (P0), potentially benchmarking the use of FTIR spectroscopy as an additional method for SOC quantification.
>
> **Proposed test:** We recommend the use of FTIR spectroscopy, with the caveats illustrated above and those discussed by Safanelli et al (2023), particularly for SIC quantification as this method performed better than acid fumigation.

As far as I understand P0 is the reference method here but also the method used in the authors research lab. It is not clear why the authors are so certain that this method is the most rigorous. For example, in Line 161-163 the authors just argue with their "expert opinion". Many labs use ball mills that are more efficient in grinding (e.g. <50um), oven drying at 105°C might cause losses of OC in some high C soils (this is only briefly touched at the end) and the pressure transducer methods requires the direct addition of acid to the soil, which can alter the organic matter (fumigation is less harsh). The authors need a reference method here to compare to but they also need to critically discuss the constrains of P0 here. It is even more important to have a good justification here given the conflict of interest that exists here between the research and the commercial lab the authors are part of at the same time.

> **RESPONSE:** We understand these concerns and so will take a new approach for this method by simply referring to it as the reference method (R) instead of P0 in the revised manuscript. Thank you for turning our attention to the fact that we should not have used the term "most rigorous" when referring to the P0. However, we will still provide references as to why we use this method at CSU throughout the manuscript. This comment also prompted us to change nomenclature of our protocol to improve clarity. The P0 method will hereby be referred to as the reference (R) and the procedural variations will be lettered according to the processing step or quantification method being tested. P1-P3 will become S1-S3 for sieving, P4 & P5 will be G1 & G2 for grinding, P6 & P7 will be D1 & D2 for drying, and P8-P10 will be Q1-Q3 for quantification. We hope this will clear up any confusion for future readers. The proposed figure for our procedural variations is included in our comments to Referee 2.
>
> We agree that ball milling generates a more homogenized, finer sample, and thus reduces analytical error, as also shown in this study. However, ball mills are typically expensive and, more importantly, have a very low throughput, making them unappealing to commercial labs. We would be curious to learn more about a high throughput ball mill if it's on the market. It is true that OC volatilization can occur at high temperatures in soils that have high OC content, but that is rarely the case in agricultural soils, and even less in those targeted for C markets. We present evidence that C volatilization did not occur in our study drying soils at 105 ºC using soils spanning a typical % SOC range found in agricultural soils. A fair point was made about acid fumigation being less harsh on the sample. However, the pressure transducer method is destructive. The sample is disposed of and does not undergo further analyses, so those transformations have no consequences. It is also worth noting that acid fumigation is often not effective at high IC levels (typically observed in deep calcareous soils) and has been reported to affect the % N, often requested with the % SOC analysis. We chose to use the pressure transducer for the reference in this study with the combination of accuracy and efficiency in mind. The acid fumigation method is more time-consuming and expensive. For standardizing methods across labs, throughput and cost is an important consideration. The authors disclosed their relationship to Cquester Analytics and, specifically to avoid any conflict which had been discussed at length

also with the funding agency of this study, we did not involve Cquester Analytics in any of this research. Throughout the manuscript, we discuss the pros and cons of all methods, and users can make up their mind as to what is most appropriate to fit their needs.

We will revise the manuscript as detailed below:

Current text (L161): Each protocol, labelled as P0-P10 (Procedure 0-10), was replicated five times per soil for all 12 soils. To our expert opinion, P0 included the most rigorous procedure at each step and all other protocols deviated from P0 for one step to enable the evaluation of the effect of each individual step on the estimation of TC, SIC and SOC concentrations.

Proposed text: Each protocol was replicated five times per soil for all 12 soils. We considered the methods used in the Soil Innovation Lab at Colorado State University (CSU) as the reference (R) where all protocols deviated from R for one step to enable the evaluation of the effect of each individual step on the estimation of TC, SIC and SOC concentrations.

Proposed text (L216): The dry combustion method (R; EA) is considered the most accurate method for total C quantification (Yeomans & Bremner, 2008) so it is often used as a reference (Leong & Tanner, 1999; Bisutti et al., 2004) against other quantification methods. SIC concentration was determined using the pressure transducer as the R method because, in our experience, it is a more efficient and cost-effective way to quantify SIC compared to acid fumigation (TC – SOC) where soil samples must be analyzed twice on the EA.

Proposed text (L490): Additionally, our finding provides evidence that volatilization of SOC is not detected in soils with < 3.6 % SOC when dried at 105 °C. The potential for SOC volatilization is a valid concern but the only study we found to test for OC volatilization prior to dry combustion corroborates our finding where there was no evidence of volatile OC loss in marine sediments dried at 110 °C and finely ground (Mills and Quinn, 1979). Additionally, we did not observe a significant effect of drying temperature on % N concentration (p=0.201; Supplemental Fig. S9). We cannot exclude the possibility that higher variability in soil C measurements would be observed in air-dry soils which had not been carefully sieved and ground, or that C volatilization would not occur in soils with higher SOC. If analysing soils with higher SOC, using a moisture correction may be preferable to oven drying the sample at 105 °C prior to EA.

Specific comments:

Line 14: Please specify what "involvement in SOC quantification for C markets" means

RESPONSE: We will specify what we mean by “involvement in SOC quantification for C markets by proposing the text (L14) read “involvement in SOC data curation used to inform C market exchanges, which could include demonstration projects, model validation and project verification activities.”

The abstract contains many details but no conclusion of the study.

RESPONSE: Thank you for pointing this out. We will add a sentence to the end of the abstract describing the conclusions as detailed below.

Proposed text (L30): We suggest that sieving to < 2 mm with a mortar & pestle or rolling pin to remove coarse materials, drying soils at 105 °C, and fine grinding soils prior to elemental analysis will improve accuracy and precision of soil C measurements. Moreover, we show promising results using FTIR spectroscopy coupled with predictive modeling for estimating % TC, % SIC, and % SOC.

Line 53: Please specify if the authors mean "quality assurance and quality control"

RESPONSE: Yes, and we will specify this is the manuscript.

Line 54: Please specify NAPT for readers that are not familiar with US organisation. This holds true for all other abbreviations that are not explained.

RESPONSE: Great suggestion. We will add more information about the North American Proficiency Testing program, and we'll go carefully through the manuscript to define all the abbreviations for readers. Specifically, we plan to add the following text.

Current text (L53): Soil testing labs can elect to participate in QA/QC certification programs that promote their data as high quality. For example, the North American Proficiency Testing (NAPT) Program is offered in the U.S., with over 130 NAPT certified soil and/or plant testing facilities. To gain certification, labs are sent soils that are similarly processed and finely ground.

Proposed text: Soil testing labs can elect to participate in quality assurance and quality control (QA/QC) certification programs that promote their data as high quality. For example, the North American Proficiency Testing (NAPT) Program of the Soil Science Society of America (SSSA) is one example of a program offered in the United States (U.S.), with over 130 NAPT certified labs. Participating labs are sent soil samples either quarterly or biannually and the data generated by each lab is subjected to a blind and double-blind statistical evaluation. Values within +/- 2.5 times the median absolute deviation (MAD) units of the median (S890 North American Proficiency Testing program oversight committee, 2020) are considered acceptable. However, labs receive soil already processed using the same methods.

Line 60: Root and rock fragments are not considered as part of the fine soil that is important for the biogeochemical processes. However, rocks and roots are still components of soils.

RESPONSE: We will add "fine" to the revised manuscript.

Line 63-65: Do the authors have any reference that commercial labs do not remove coarse fragments. To my experience, research labs apply sieving and in general same sample preparation for agricultural and non-agricultural soils. Also, soil inventories prepare the fine soil prior to C measurements.

RESPONSE: Unfortunately, we could not find a published study demonstrating that commercial labs do not remove coarse fragments. However, we believe we have support for this claim in the current manuscript. We presented results from a preliminary survey that showed over 70 % of the service labs surveyed use a mechanical flail grinder for the initial sieving step (L74; Supplemental Table S1). We also showed in Supplemental Table S2 that 5 of 8 labs used in our blind comparison sieve with a mechanical grinder and only 1 of 8 fine grind the sample beyond the 2 mm sieve (or in one case the 1 mm sieve). Because the whole bulk sample is poured into the grinder prior to falling over the 2 mm screen, it's safe to assume that coarse material is ground before being

> removed. In the case that coarse fragments are picked out of the sample after a pass through the mechanical grinder, there still may be some that goes through the 2 mm screen initially.

Line 65-67: It is not clear to me why regenerative agriculture results in more coarse fragments in deeper soil. Also, the authors refer here rather to conservational land management rather that regenerative land management, which is a very broad and not well-defined term.

> **RESPONSE:** We agree that regenerative agriculture is a broad term and can involve many different types of management. We will link regenerative agriculture and deep-rooted perennials crops better as detailed below.
>
> **Original text (L65):** Compared to conventionally managed agricultural fields, coarse materials are more abundant in deeper soils in regenerative agricultural lands that include cover or perennial crops and grasses, thus it's important to consider how coarse materials in these soils may affect C estimation.
>
> **Proposed text:** Compared to conventionally managed agricultural fields, agricultural lands managed using a regenerative practice, like the addition of certain perennial crops (i.e., alfalfa), typically have more coarse materials deeper in the soil profile as more root biomass is incorporated at depth (Fan et al., 2016). Thus, it's important to consider how coarse materials in these soils may affect C estimation.

Line 77: Also here, the authors should be specific since it is considered as "fine soil"

> **RESPONSE:** We will add "fine".

Line 97: The authors should specify if near-infrared of mid-infrared regions.

> **RESPONSE**: Thank you for pointing this out. We will add "mid-".

Also, such approaches require a well-trained model based on large enough soil spectral library. This is a critical step for the quantification of soil C using chemometric approaches. Therefore, it follows a different concept compared to the other more direct methods.

> **RESPONSE:** Please refer to our response above.

Line 121: I would rather expect that the dual homogenisation by sieving to 8 followed by 2 mm would result in lower variability.

> **RESPONSE:** Yes, that is a good point, but we will keep our original hypothesis.

Table 1: is pH, %SOC and %SIC are measured with analytical replicates? The authors should add errors to the values.

> **RESPONSE:** Yes, % SIC and % SOC have replicates (n=5). We will add the standard deviation to % SOC, and % SIC and add a column for % TC with standard deviation reported.

Table 1 caption: How was pH measured and what are the texture classes applied? It is not clear what "Colorado State University following procedure P0" is. Please provide details of refer to Table 2 here.

RESPONSE: We agree so will provide more details for the pH and texture methods used and refer to Table 2 in the caption for the reference methods.

Proposed text (L141): The pH was determined using a 1:1 ratio of soil to deionized water. Texture was determined after shaking 40 g of soil in 5 % sodium hexametaphosphate solution for 18 hours, wet sieving sand > 53 µm, and using a hydrometer to determine silt and clay content. Texture classes were defined according to the soil texture calculator created by the Natural Resources Conservation Service U.S. Department of Agriculture.

Line 218-219: This is not very precise. It is not clear which model and was used and on which data it is trained.

Line 223-224: This is most likely attributed to the fact that the used model for the prediction based on the KSSL is developed with samples of similar degree of grinding. In the cited paper, Sanderman et al (2023) conclude that the model trained on fine milled samples was not well transferable on the coarser samples.

Therefore, the authors used a model that was trained for a certain milling. This makes this testing of grinding here not very useful. in comparison, Sanderman et al (2023) developed separate models for roughly 2400 samples of the KSSL. They conclude that a model that was trained with coarse samples and predicted coarse samples was performing similar to a model that was trained with fine soils and predicted fine soils. However, the transfer of models was not satisfying.

RESPONSE: We agree and very much appreciate this insight. Please refer to our response to the related general comment made above.

Line 231: Before it was mentioned that acid fumigation was only performed for P9. Here it reads like every sample was fumigated. Please clarify.

RESPONSE: We will clarify that acid fumigation was only used for the P9 (future Q2) procedure.

Line 241-243: How were CO2 and H2O interferences corrected?

RESPONSE: Thank you for this reminder. We will include a sentence describing how the CO_2 and H_2O interferences were corrected as detailed below.

Proposed text (L243): A background of gold was scanned before every sample to correct for potential fluctuations and interference of CO_2 and H_2O.

Line 245-248: The authors should add more details regarding the predictions. Seybold et al (2019) developed PLSR based on the NSSC-KSSL. is this also used here? What do the authors mean with "respective geographical region"? Were the models local? Was there any spectral pre-processing like re-sampling, filtering, normalization or bassline correction?

RESPONSE: These are valid questions that we will address as detailed below. We will also include a supplemental table with each model summary.

Current text (L245): For predicting % TC, % SIC, and % SOC, spectra were trimmed from 4000 to 600 cm^{-1}. Models were built separately in the OPUS software (OPUS version 8.5, Bruker Optik GmbH 2020) for % TC, % SIC, and % SOC for each soil's respective geographical region using the USDA NRCS National Soil Survey Center-Kellogg Soil Survey Laboratory (NSSC-KSSL) spectral library (Seybold et al., 2019).

Proposed text: For predicting % TC, % SIC, and % SOC, calibration models were built using the USDA NRCS National Soil Survey Center-Kellogg Soil Survey (NSSC-KSSL) spectral library coupled with partial least squares regression in the OPUS software (OPUS version 8.5, Bruker Optik GmbH 2020) as described in detail by Seybold et al. (2019). The calibration models were developed separately by soil property and geographical region (i.e., the state of Colorado (CO) was used as the boundary for making predictions with soils collected in CO). Spectra were trimmed to the mid-infrared region from 4000 to 600 cm^{-1} and calibration spectra were mean centered with redundancies removed using principal component analysis and outliers removed based on ANOVA of residuals in OPUS. Details for the geographical boundaries, spectral pre-processing, R², and root mean square error of prediction for each model can be found in Supplemental Table S3.

Table S3: Summary for each model built using the United States Department of Agriculture Natural Resources Conservation Service National Soil Survey Center-Kellogg Soil Survey coupled with partial least squares regression in OPUS (OPUS version 8.5, Bruker Optik GmbH 2020) describing the soil property of interest for prediction, spectral library boundaries, spectral pre-processing for model optimization, and the validation model R² and root mean square error of prediction (RMSEP).

Soil Property	Area Name	Spectral pre-processing	R²	RMSEP
Total Carbon	Colorado	First derivative + Vector normalization (SNV)	0.9506	0.433
Total Carbon	Wyoming	First derivative + Vector normalization (SNV)	0.9543	0.442
Total Carbon	Iowa or Nebraska	First derivative + Vector normalization (SNV)	0.9625	0.309
Total Carbon	Kansas	First derivative + Vector normalization (SNV)	0.984	0.44
Inorganic Carbon	Colorado	First derivative + Straight line subtraction	0.9899	0.834
Inorganic Carbon	Wyoming	First derivative	0.9925	0.761
Inorganic Carbon	Iowa or Nebraska	First derivative + MSC	0.9768	1.56
Inorganic Carbon	Kansas	First derivative + Straight line subtraction	0.9959	0.798
Organic Carbon	Colorado	First derivative + Vector normalization (SNV)	0.963	0.398
Organic Carbon	Wyoming	First derivative + Vector normalization (SNV)	0.9604	0.318
Organic Carbon	Iowa or Nebraska	First derivative + Vector normalization (SNV)	0.9553	0.318
Organic Carbon	Kansas	First derivative + MSC	0.9564	0.257

Line 270: "External service labs provided values for % TC, % SIC, and % SOC." can be removed.

RESPONSE: We will remove this sentence.

Line 271: Looking at Table S3, it seems not fair to just select the extremes here. Most differences are rather lower. It is hard to tell from the table. Maybe boxplots per soil with different symbols for the labs would be easier to read. Anyway, the authors should also mention the range of differences and not only the extremes.

RESPONSE: This is a good suggestion, thank you. We only reported on the extremes in the current version because that's a big take away with this dataset, but it's true that, in some cases, the distribution is much tighter. We will add to the text to include more results for the blind comparison as detailed below and update Fig. 1 to present the data more clearly. In the revised manuscript Fig. 1 will be Fig. 2 as proposed below.

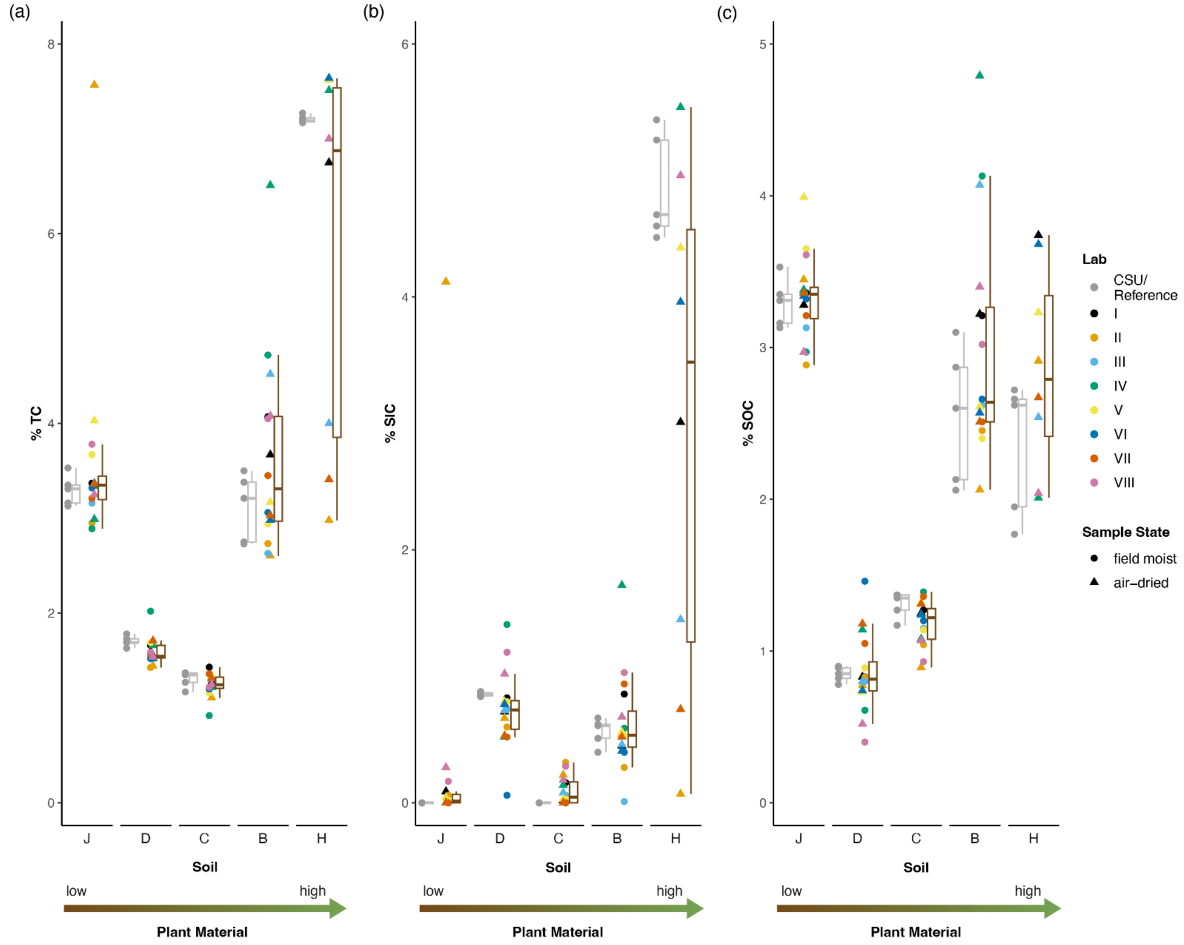

Figure 2: The distribution of total carbon (TC; panel a) soil inorganic carbon (SIC; panel b) and soil organic carbon (SOC; panel c) concentrations from eight service soil testing laboratories and Colorado State University (CSU). Box plots report the median, first and third quartiles for values from all soils (field moist and air-dried) analyzed at service soil testing laboratories (brown boxplot) and CSU (grey boxplot; n=5). Whiskers extend to the upper and lower data point that are within 1.5 times the interquartile range. For soils B, C, D, and J, two samples were sent to each external

lab, one air-dried and one field moist (n=16). One sample from soil H was sent to each lab (n=8). Refer to Table 1 for a description of the soils, Figure 1 for Reference (CSU) methods, and Supplemental Table S2 for external service soil testing laboratory methods.

Proposed text (L274): However, in some cases, labs reported the same or similar values either air-dried or field moist. For example, Lab VII, reports no difference in % SOC while Lab I only detected a 0.01 % difference between the air-dried and field moist samples sent from soil B. Lab VI reported differences of < 0.1% TC for all soils sent while Lab I reported differences in SOC of < 0.1 % for all soils.

Line 284: Yes, it is an astonishing range of measured values between labs. It is also surprising that the reference measurement (CSU lab) shows a large variability of soil B and H. These are two soils with high pH. I wonder if this could be an effect of the carbonate removal. What is your explanation for large differences between the five analytical replicates? Additional, the external labs did not measure in replicates?

RESPONSE: Yes, we realize that there is notable variability within the CSU lab for these two soils. These two soils have the highest variability by far, so we wanted to send them to the external labs for comparison. However, we will note that when the distribution of CSU's 5 reps is compared to the distribution across the external labs, the external labs (viewing each data point like a replicate) have a larger distribution. Given that soils were homogenized for subsamples using the same method for CSU and for the external labs, we attribute this to differences in soil processing used across labs. We will add our explanation for the variability and compare it to the variability across the external labs as detailed below:

Revised text (L403): However, we observed notable variability across the CSU reference for soils B and H. We speculate that the higher variability in % SOC in soil H is due to the measured values for % SIC using the pressure transducer since % TC variability is very low and % SOC is calculated by % TC - % SIC for the CSU (R) method. The presence of substantial amounts of fine root and some SIC in soil B (irrigated pasture) is most likely the reason for high variability using the CSU R method. By sending external labs the two soils with the highest variability, we were able to confirm the expectation that high % SIC and high fine root material contribute to higher variability in SOC data. Additionally, we can attribute higher variability to differences in processing methods as, despite a notable distribution in soils B and H from CSU, the distribution across all external labs is much larger (Fig. 2).

We regret that we did not have enough soil to send 5 replicates of field moist and air-dried soils to each external lab. That was an unfortunate oversight. We describe this in the current manuscript L150.

Line 288: Why are the no coarse materials at all in soil L for the P3?

RESPONSE: Good observation. There was no coarse material collected from any of the replicates using the S3 procedure, meaning that it was all ground into the sample and considered fine soil.

Line 305-307: This relationship seems to be mainly driven by the on P3 point at 0.8 difference in plant material and 1 STD %SOC (right corner). This is in general a very weak correlation and might not add much when the one point is considered as an outlier.

RESPONSE: We discussed this among the co-authors as a potential issue prior to the preprint so will take your point of view into consideration as well and remove Fig. 3 in addition to any text referring to the figure. Thank you for your input.

Line 323-324: Do the author mean a relationship to SOC, similar to the plant material in Fig. 3?

RESPONSE: Yes, we will clarify that in the revised manuscript.

Figure 4 and results section: This paper is mainly about errors that are important for the SOC because this will be of interest for the C markets. Therefore, I wonder if Figure S3 with the SOC differences between soils and methods should be the main figure in the manuscript and the current Figure 4 could more to the SI. This might need a restructuring of the section as well.

RESPONSE: Yes, that's true. We will replace Figure 4 with the proposed figure below in the main text and revise the text accordingly.

NEW FIGURE:

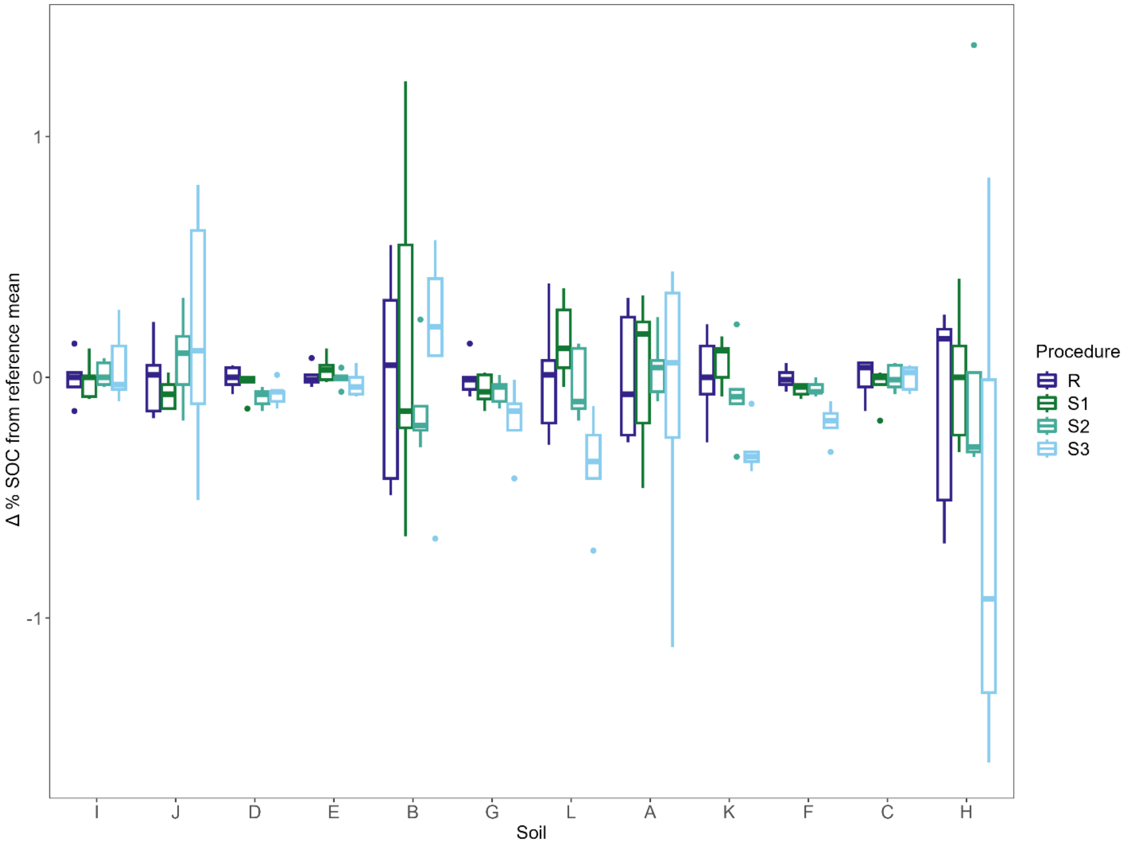

Figure 4: The difference (Δ) in % soil organic carbon (SOC) compared to the reference (R) mean value for all sieving procedures including R as described in Figure 1. Box plots report the median, first and third quartiles. Whiskers extend to the upper and lower data point that are within 1.5 times the interquartile range. Letters indicate the different soils, as described in Table 1, which are arranged on the x-axis by proportion of rock material removed with the R sieving procedure.

Line 339: Significances are shown in Table S6?

> **RESPONSE:** Yes, the referee is correct. We will add that to the text for readers.

Figure 6: X axis label, colour and legend are redundant.

> **RESPONSE:** We will improve this figure in the revised manuscript.

Line 396-397: Here the focus is on SOC for the C market.

> **RESPONSE:** The referee is correct. We will clarify SOC instead of C, in the revised manuscript.

Line 398-399: Please see my comment regarding the variability on CSU lab for some soils. This is also concerning. Here it would be good to have replicates from the individual labs.

> **RESPONSE:** Please refer to our response above.

Line 460: This would be a very interesting aspect of the manuscript. The C market needs stocks of C and not concentrations alone. Therefore, the effect of removed or not removed coarse fragments would be most significant. Even the calculation of SOC stocks includes large uncertainties and this would add up with the method uses (e.g. Poeplau et al. 2017). The authors do back on the envelope calculations later in the implications section. Would it be possible to discuss the stock effects even more by estimating the stock differences here for all methods using the soils bulk densities?

> **RESPONSE:** We appreciate the enthusiasm about this aspect we added to the manuscript. While we would love to honor this suggestion, we were unable to calculate bulk density because we did not get an accurate volume of soil when we collected it by shovel. For that reason, we chose to assume a bulk density of 1 g/ cm^3.

Line 465-468: Yes, plant material would be low in mass but might be important in volume and thus could have an impact of stocks as well.

> **RESPONSE:** That's a good point but we do not consider the volume of roots in our bulk density calculations as described in Poeplau et al. (2017). Only the mass is considered.

Line 567: This should be Fig. S9

> **RESPONSE:** Thanks, we will correct the figure number.