Reply to RC2: 'Comment on egusphere-2026-1255', Anonymous Referee #2, 26 May 2026

We sincerely thank the Referee for the valuable and constructive comments. We respond point by point below.

Major comments:

The central research question is based around whether rapid or field-deployable tools are interchangeable with standard laboratory methods. The scientific question is interesting and important, although ultimately unanswerable using the experimental design and statistical analyses.

The methods show a sampling design of 6 subplots (3 plots × 2 management treatments) with a single point per subplot sampled at two depths sampled at 11 time intervals. This design has three nested sources of dependence: subplot (n = 6), time point (n = 11), and depth (n = 2). The real replication for the management treatment contrast is n = 3 subplots per treatment. The Spearman rank correlations and RDA appear to treat all 132 observations as independent, which is pseudoreplication. The consequences are that the Spearman rho confidence intervals are narrower than they should be, and p-values are inflated, possibly by orders of magnitude; RDA permutation tests are biased because permutations assume independence among the observations; and rank concordance between on-site and laboratory methods may largely reflect a combination of the depth gradient and within-subplot temporal autocorrelation, which are not accounted for. For the method-comparison Spearman correlations, a more defensible approach would be to aggregate observations to the subplot level (or to compute within-subplot rank correlations and test the distribution across subplots) before computing inter-method correlation statistics. For the RDA, time should be included as an adjusted covariate and permutations stratified by subplot. Without restructured analyses that account for the nested and repeated-measures design, the manuscript cannot support the asserted rank concordance between on-site and laboratory methods.

Throughout the manuscript the authors describe their analyses as testing the comparability and agreement between on-site and laboratory methods, but the statistical framework applied (Spearman's rank correlation) does not test comparability, agreement, or sensitivity, but instead whether two methods order samples consistently. It is uninformative about whether one method could substitute for another. The visual presentation compounds this problem, as linear regression lines with confidence intervals are plotted on scatter plots showing one method against the other, while reporting Spearman's rho on the same plots. The authors should restrict their claims to rank concordance, removing all "agreement" and "comparability" to "the methods rank samples consistently" or emphasizing the relative ordering nature of the comparison. The rank correlation comparison is still useful and important to the literature.

ANSWER: We thank the Referee for this critical observation regarding the nested structure of our data. For the Spearman rank correlations, we will restructure the analysis following the Referee's suggestion by computing within-subplot rank correlations between on-site and laboratory methods across time points and depths, and testing whether the distribution of rho values across subplots is consistently positive. For the RDA, we will include time as an adjusted covariate to partial out temporal variance and will stratify

permutations by subplot. We acknowledge that these restructured analyses may yield more conservative results, and we will revise our interpretations accordingly.

We also agree that Spearman's rank correlation tests rank concordance, not agreement or substitutability between methods. We will revise all language throughout the manuscript, replacing "agreement" and "comparability" with "rank concordance" or "consistent sample ordering." We will also correct the figures by removing the linear regression lines and confidence bands, replacing them with a 1:1 identity line, which is the appropriate visual reference for a rank correlation comparison.

Minor comments:

28-30 One of the citations for this assertion is a "viewpoint" and should be removed

ANSWER: The Referee is right, we will remove this citation.

40-41 Slakes is not a field-based measurement, as aggregates must be air-dried prior to analysis

ANSWER: We agree with the Referee and we will correct the manuscript by removing the term 'field-based'. The focus of the study is on analytical speed and infrastructure requirements rather than strict field deployment.

47-53 Assessing the performance of rapid laboratory measurements relative to traditional laboratory measurements through numerical comparison alone makes two assumptions: 1. That the methods are measuring similar properties using similar methodologies and thus have a basis for numerical comparison; 2. That traditional laboratory measurements are "better", i.e. more sensitive or precise. I am not convinced that either assumption is met, and the manuscript would benefit greatly from addressing these assumptions. There are other studies that quantitatively evaluate these methods based on their sensitivity to management (e.g., Rieke et al., 2023)

ANSWER: We agree that these conceptual assumptions deserve direct attention in the manuscript. Regarding the first assumption, we acknowledge that the methods compared in this study do not always share similar methodologies, even when targeting the same soil property. As already discussed in the manuscript for respiration and enzymatic activities, some comparisons involve methods that measure different operational expressions of the same underlying biological process rather than equivalent analytical procedures. We will revise the framing in lines 47-53 to make this distinction explicit, clarifying that our comparisons assess rank concordance between methods targeting the same soil health indicator, rather than numerical equivalence between analytically interchangeable procedures.

Regarding the second assumption, we fully agree that traditional laboratory methods should not be treated as an unconditional gold standard. Their widespread use reflects

historical adoption rather than demonstrated superiority in sensitivity or ecological relevance. We will clarify this framing in the Introduction.

64-67 The LUCAS framework was designed for landscape-scale soil chemistry and physical property monitoring, such as total C/N, pH, and soil texture, and are not appropriate for biological assessments, since typically the 20-50 cm depth is not very biologically active.

ANSWER: We agree that the LUCAS framework was originally developed primarily for large-scale monitoring of soil physicochemical properties. However, biological properties have also been incorporated into the LUCAS monitoring scheme using the same sampling depths (Orgiazzi et al., 2017, https://doi.org/10.1111/ejss.12499). Although biological activity is generally lower in subsoil layers, we consider that assessing biological indicators at depth remains relevant, as deeper soil horizons contribute to important ecosystem functions and may respond differently to management and environmental drivers.

76-83 Were aggregates air-dried prior to running Slakes as required by the methods? This would seem to mean it is not a field-based test. Additionally, only ~9 soil aggregates 3-10 mm are typically required for Slakes as implemented commercially, which would translate to 1-5 g, maybe enough for one Eijkelkamp run. What was the actual mass of soil used to for Eijkelkamp, and how many replicates were run for each test? The aggregate sizes are likely different for both tests, with Eijkelkamp measuring aggregate stability on 1-2 mm aggregates and Slakes measuring aggregate stability on 3-10 mm aggregates. This holds implications for comparability. Finally, Slakes was validated on aggregates collected from 0-6 cm in depth, and has not been tested for other depths, please acknowledge or address

ANSWER: The objective of this study was not to compare identical analytical procedures, but rather to evaluate whether rapid methods and standard laboratory methods capture similar patterns and trends. Nevertheless, for the Eijkelkamp wet-sieving procedure we adapted the standard protocol to maximize comparability with SLAKES, using aggregates of similar size. The measurements were performed in triplicate. We will clarify the aggregate size range and replication scheme in the revised Methods section.

We will also acknowledge in the manuscript that SLAKES was originally validated using 0-6 cm depth aggregates. However, we note that SLAKES has subsequently been applied in studies evaluating deeper soil layers (e.g., Flynn et al., 2025, https://doi.org/10.1002/saj2.20012; Adetsu et al., 2024, https://doi.org/10.1002/saj2.20674).

84-89 The comparison between ISO 16072 bench respiration and in-situ IRGA flux is not methodologically defensible, as the two methods differ in soil disturbance (air-drying, sieving, rewetting), environmental conditions, and what type of respiration is measured. Due to the loss of microbial biomass through air-drying, Birch-effect flush of CO2, presence or absence of autotrophic root respiration, and difference in ambient soil moisture, the alkali trap method will likely overestimate low fluxes and underestimate high

fluxes (Yim et al. 2002; Jensen et al. 1996). So many sources of method bias cannot support the inferences the authors seek to draw; I recommend either removing this comparison or replacing the disturbed soil protocol with a minimally disturbed intact-core incubation

ANSWER: We agree that substantial methodological differences exist between in-situ IRGA measurements and laboratory respiration assays conducted according to ISO 16072. Some of these differences are already acknowledged in the Discussion, and following similar suggestions from Referee 1, we will further expand this section to more explicitly address the inherent decoupling between in-situ and laboratory respiration measurements. However, if the Referee considers that the methodological differences make this comparison inappropriate, we would be willing to remove the respiration comparison from the manuscript.

90-96 This section compares a biomass-based ratio with a relative abundance ratio (DNA copy count), which makes comparison difficult; additionally, ITS is more accepted for fungal community profiling, please justify

ANSWER: There is a misunderstanding regarding the comparisons performed in this study. The microBIOMETER® provides two different outputs: an estimate of microbial biomass and an estimate of the fungal-to-bacterial ratio. Accordingly, we compared microbial biomass estimated by microBIOMETER® with microbial biomass carbon determined by the chloroform fumigation–extraction method, while the F:B ratio reported by microBIOMETER® was compared with the F:B ratio derived from qPCR quantification of fungal and bacterial abundances. We will revise the text to clarify this.

Regarding fungal quantification, ITS but also 18S rRNA markers have been widely used in the literature (e.g., Banos et al., 2018; https://doi.org/10.1186/s12866-018-1331-4). While ITS generally provides higher fungal specificity and is often preferred for community profiling, the 18S rRNA gene offers a more conserved target with a relatively constant amplicon length across taxa, which can improve qPCR efficiency.

102-112 The fact that Spearman's rank correlation was used to compare methods needs to be highlighted much more in the introduction; the goal of this paper is not numerical comparison, but comparison of the ordering of treatment pairs. Still, potential confounding remains and much more caution should be used in interpreting results

ANSWER: We will add explicit framing in the Introduction clarifying that the objective of this study was to evaluate rank concordance between methods rather than numerical equivalence. We will also revise our interpretative language throughout the text to reflect this.

120 The linear regression line is not appropriate, as it implies that y has measurement error but x does not, that the relationship is linear, and that errors are normally distributed; the confidence interval is for the conditional mean of y given x, not the prediction or

agreement interval; a 1:1 line should be plotted if the goal is method comparison; and the inclusion of the Spearman values is inconsistent with the linear regression

ANSWER: The Referee is correct. As stated above, we will correct the figures by removing the linear regression lines and confidence bands, replacing them with a 1:1 identity line.

134 There is a substantial body of literature that addresses all of the differences between field-measured respiration and laboratory-measured respiration, please research, consult and amend

ANSWER: We thank the Referee for this suggestion and agree that there is an extensive body of literature describing the conceptual and methodological differences between field-measured and laboratory-measured soil respiration. If the respiration comparison is retained in the manuscript, we will expand the revised manuscript to better reflect the existing knowledge on this topic. In particular, we will incorporate references discussing the effects of soil disturbance, incubation conditions, moisture and temperature controls (e.g., Ananyeva et al., 2020, https://doi.org/10.1134/S106422932010004X; Patel et al, 2022, 10.1088/1748-9326/ac9aca; Li et al., 2025, https://doi.org/10.5194/bg-22-2691-2025).

140-142 This is a good point and should be expanded

ANSWER: We agree that this point deserves further development. As noted in our responses above, if the respiration comparison is retained in the manuscript, we will substantially expand this section of the Discussion to better explain the conceptual differences between field and laboratory respiration measurements, the distinct ecological information provided by each approach, and the implications for soil health assessment and interpretation of results.

167-175 These should be amended to address the feedback provided in major comments

ANSWER: We agree that the Conclusions require revision to align with the methodological corrections and reframing addressed in the major comments. Specifically, we will replace "agreement" and "closely matched" with language reflecting rank concordance rather than numerical equivalence. Besides, we will add appropriate caution regarding the pseudoreplication issue and the nested structure of the data, tempering claims of concordance where the restructured analyses may yield more conservative results.