Dear Referees,

Thank you very much for your valuable feedback, suggestions and inputs. This will undoubtedly contribute to the enhancement of manuscript's quality. Below you will find a general outlook describing how we plan to address the comments that you provided.

Introduction, study design and discussion

- The flow between the paragraphs of the introduction will be improved. The C isotopes will be better introduced and research questions will be merged into one clear question: "Down to which depth do the different organic residues affect SOC stocks".
- The subsoil will be more clearly defined in the introduction, with an explanation for the 30 cm theshold between top- and subsoil.
- The specific recommendations to change phrasing will be considered and the text will be improved accordingly.
- Overall, some more detail will be added on the pH, CEC and C isotopes and on how these parameters are affected by organic residues in the introduction and in the discussion.
- Our study shows that organic residues affect SOC below the commonly studied soil depth of 30 cm. However, it remains unclear whether increased subsoil OC can impact crop productivity. This aspect will be emphasized more, as it warrants further investigation.
- It will be clarified that the results from Laub et al. (2023b) were based on a different sampling campaign than the samples used in our study.
- In Section 2.1.1, we will add a statement clarifying that "*all results indicating higher SOC for a given treatment in our study should be interpreted as losing less carbon, as Laub et al. (2023b) indicated that all treatments have been consistently losing SOC since the initiation of the experiment.*".
- The explanation of OC stocks calculations will be clarified.
- The limitations of our results will be further highlighted, and their interpretation will be presented accordingly. The reasons behind the choices of the applied statistical analyses will be provided. For example, we will put more emphasis on the fact that merging the ±N variant of the organic residues treatments has its limitation as the long-term field trial was designed to study the different impact of these treatments. However, we will highlight the fact that without merging the ±N variant the statistical power (which was already low for most analyses) would be much lower and inflate the risk of type II error.

Improvement of data anlysis:

- Following the feedback of the reviewers, the linear mixed model will be adjusted as follow:
  To avoid overfitting and ensure the robustness of our findings, we will exclude interactions in the model. This approach will simplify the model, focusing on the main effects and accounting for variability through random effects, thereby reducing the risk of capturing noise rather than true underlying patterns. We will evaluate the robustness of our model by fitting six different linear mixed-effects models using the lmer function from the lme4 package in R. Each model will include OC stocks as the response variable, the sampling block as a random effect, and different combinations of fixed effect predictor variables:
  - model 1 (depth, organic residues, mineral fertilizer, silt content, silt content)
  - model 2 (depth, silt content, silt content)
  - model 3 (depth, organic residues)
  - model 4 (depth, mineral fertilizer)
  - model 5 (organic fertilizer)
  - model 6 (mineral fertilizer)

  We will then compare these models with an ANOVA to determine the best-fitting model based on AIC, BIC, log-likelihood, and deviance values. Only the reuslts of the best fitting model will be presented in the manuscript and additional information will be available in the suppementary information. To further test the effect of the organic residue treatment in the deeper soil layer, the same analysis will be performed on a subset containing only subsoil layers.

- We will justify the use of a *t*-test over other statistical tests and explain why we are convinced that the risk of a type I error is small compared to the risk of Type II error. We acknowledge the concern of reviewer 2 regarding type I errors with multiple *t*-tests. However, our study is hypothesis-driven, and we observe a clear trend: significant differences observed in the top layer diminish with depth. As we analyze deeper soil layers, the observed differences between OR treatments and the control decrease. We perform this analysis layer by layer to determine down to which depth these statistical differences are observed. We are not randomly testing a large number of treatmnents against each other (in which case the multiple *t*-tests would indeed increase the risk of type I errors). Given the limited data (due to sampling in a long-term field trial) and increasing variability with depth, we chose *t*-tests over an ANOVA because an ANOVA would be even more affected by the limited statistical power. When statistical significance was not achieved, the power was below 80%, supporting our choice. Therefore, we argue that the typical type I error inflation does not apply in our context and that we are more likely to experience type II error.