This paper reports soil organic carbon (SOC) concentrations and stocks from a long-term residue, manure, and fertilization experiment. The authors discuss the changes in SOC between residue and manure treatments and over time in the context of soil carbon saturation and the '4 per mille' recommendation, providing the hypothesis that crop residue incorporation and poultry manure addition may be effective for fulfilling goals of '4 per mille'. While most of the data are novel and the general concept of reporting SOC levels in a long-term field experiment is sound, I have multiple concerns surrounding almost all aspects the of the sample collection, discussion of experimental design, data reporting, analytical methods, interpretations of results, and writing style.

We thank the reviewer for appreciating the amount of work and people that are behind this type of long-term experiments and for pointing out these issues that were promptly considered for manuscript improvement.

Specific comments:

Lack of clarity about number of cores taken per experimental plot for SOC analysis. The methods section is written in a way that it is not entirely clear how many soil cores were taken per plot, but it seems possible that there were 60 'disturbed' cores taken in total, one for each experimental plot? If so, this is a strong limitation of the study. SOC is highly spatially heterogenous, so if one core per plot was sampled, even with 4 plot-level replicates of each treatment, this coring strategy would introduce likelihood that true differences in SOC levels due to treatment were not clearly observed and that reported differences reflect a large element of spatial heterogeneity across the field trial rather than showing treatment effects. If one core per plot was taken, while the sampling strategy cannot be altered at this stage, I would strongly suggest to the authors to clearly report the number of cores taken per plot for SOC analysis (e.g., "We sampled one core per plot for SOC analysis, and divided it into two depths") rather than only total number of cores, and to justify this approach, and to acknowledge in the discussion section the limitations introduced into the detection of treatment differences by this likely under-sampling of the field trial.

- We got the point of the reviewer, and we agree. We will modify the entire sampling section clarifying the sampling procedure. In brief, the soil sampling involved two different samplings. 1) 60 undisturbed soil cores (7 cm diameter, 60 cm height) were collected from the middle of each plot using and hydraulic sampler, cut into distinct layers (0-30 and 30-60 cm) and oven-dried for bulk density determination. 2) Disturbed soil samples were collected from 5 positions inside each plot in the 0-30 cm and 30-60 cm layers using a hand-push auger. Afterwards, the 5 sub-samples per plot per depth were mixed to form 120 (60 plots x 2 depths) soil samples.
- Confounding of tillage with residue retention in experimental design. The experimental design consisted of two treatments of residue incorporation (one with manure) and another with residue removal, at all levels of N fertilization, where only the treatment of residue removal was not disturbed, thus presenting

a confounding effect of tillage with residue. Tillage can affect amount and distribution of SOC. While this long-term treatment is not under the control of the authors, the rationale for the experimental design, its drawbacks, and the ambiguity it introduces into interpretation of the results should nevertheless be carefully discussed.

- The tillage practices were the same across the entire experimental area meaning that all treatments, including residue removal, received the same tillage operations. We will clarify this aspect in the text by stating that "the tillage operations were the same in all treatments, consisting in soil ploughing followed by rotary harrowing".
- Use of different analytical methods to measure SOC over time. The authors report pre-existing data from a time series of SOC sampling at the site, where earlier samples were assessed with the dichromate oxidation method and later samples were analyzed with a flash combustion method. Different analysis methods for SOC return different results for the same soils (Roper et al. 2019 https://doi.org/10.2136/sssaj2018.03.0105); while the authors performed a methods comparison with this in mind, the results of this comparison are not reported (paragraph line 115). I suggest that the authors report summary statistics and show data, possibly in supplementary materials, of their methods comparison tests and also incorporate a discussion of how the interpretation of time series data may be affected by variability and especially bias introduced by the different SOC analysis methods.

- We got the point of the review, and we agree in considering the analytical methods a thorny subject especially when dealing with long-term experiments. She/he will probably already know that all the research groups running LTE spend a lot of effort in keeping the data comparable across different years. Up to 1994, the dichromate oxidation method was used to analyse SOC concentration. In 2006 flash combustion (Elemental Analyzer) was introduced, and, in this study, we used the novel high-temperature catalytic combustion according to DIN19539. For the first analytical technique change, we obtained a slope of 0.99025 while 0.9221 for the second method change. Here below you can see the related summary statistics of those regressions. We agree that this is a central issue, and we can add the regression in the supplementary material of the final version of the paper. Moreover, in the discussion section, we can add a paragraph speculating on how different analytical techniques might have affected our results.

<table>
<tr>
<td>Flash combustion vs dichromate oxidation

Call:
lm(formula = Conc ~ 0 + WB)

Residuals:
Min 1Q Median 3Q Max
-0.45384 -0.17619 -0.07775 0.06149 1.28087</td>
<td>High-temperature catalytic combustion vs flash combustion

Call:
lm(formula = Skalar ~ 0 + CNS)

Residuals:
Min 1Q Median 3Q Max
-0.113312 -0.030874 -0.004654 0.040663 0.120899</td>
</tr>
</table>

<table>
<tr>
<td>Coefficients:
Estimate Std. Error t value Pr(>|t|)
WB 0.99025 0.03489 28.38 <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3346 on 35 degrees of freedom
Multiple R-squared: 0.9584, Adjusted R-squared: 0.9572
F-statistic: 805.4 on 1 and 35 DF, p-value: < 2.2e-16</td>
<td>Coefficients:
Estimate Std. Error t value Pr(>|t|)
CNS 0.922089 0.005947 155 <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05313 on 74 degrees of freedom
Multiple R-squared: 0.9969,
F-statistic: 2.404e+04 on 1 and 74 DF, p-value: < 2.2e-16</td>
</tr>
</table>

- Non-reporting of effect of N fertilization on SOC stocks. While the authors only report SOC levels per residue and manure addition treatment, there were five levels of N fertilization also sampled. In Figures 1-3, which N fertilization levels are represented? Or, are N fertilization levels averages across residue and manure treatments? Please specify in the methods and in the figure legends.
- We did not focus much on N since only a few significances were evidenced. Anyway, we understand the comments of the reviewer and we will add more information regarding the N-level effect on SOC. In Figg 1-3 the average of all N levels is reported. We will improve the figure legends, as suggested. Moreover, we will add a new figure showing the effect of N on SOC concentration, where significant.
- Non-reporting of C input estimates. Although a calculation of C inputs is discussed in the methods section and reported in the abstract, the data on C inputs across treatments are not currently reported in any table, or figure, or in the results section. Please add these results.
- We appreciated the reviewer's comment which will be promptly considered for MS improvement by adding text, figures/table on C input.
- Absence of results section corresponding to Figure 3 (SOC change over time). The authors present Figure 3 in the discussion section, and these data are discussed in the abstract and conclusions, but currently there is not a results section corresponding to this figure. I suggest the authors either expand existing results sections to include results reporting for this figure, or create a new results section.
- We got the point of the reviewer, and we will move Fig. 3 to the results section and will expand this part by also adding the corresponding text.
- Overall lack of legibility of data visualization. Particularly in Figure 2, the shadings used for different depths can't be discerned from the legend. The caption in Figure 2 doesn't specify if post-hoc comparisons represented by letters are across treatments within the same depth; this seems likely but could be clarified (same in Figure 1). In Figure 1, larger font size for axis labels would improve readability.

- We will revise the Figure by increasing the readability, as suggested.
- Claim of testing the soil carbon saturation concept: the authors use previous literature to calculate the expected maximum of soil organic carbon in mineral-associated form (MAOC) based on soil texture at the experimental site, and reasonably point out that the SOC levels observed were below the expected theoretical maximum, and therefore far from saturation. While their conclusion is probably sound, there are several conceptual and analytical flaws with the work, and a more nuanced approach and discussion would improve the rigor of the interpretation and claims. First, the concept of soil carbon saturation applies specifically to the mineral-associated carbon (MAOC), isolated through soil disturbance and size or density cutoffs. However, the authors measure and report only total soil organic carbon, which also includes particulate organic carbon (POC) that is not theorized to be controlled by saturation limits. Since the SOC (=MAOC + POC) reported is still below the theoretical maximum of MAOC based on soil texture, the claim that the soils were below saturation (based on this theory and method accounting) is still accurate, but it needs to be acknowledged the implications of the presence of POC in the sample and how this affects the saturation estimate. Second, the authors use older references for their calculation of MAOC at saturation; why use these rather than larger and more recent datasets? E.g., Feng et al. 2013 1007/s10533-011-9679-7; Georgiou et al. 2022 10.1038/s41467-022-31540-9.
- We got the point of the reviewer, and we agree. Regarding the first raised point we will specify that the saturation concept applies to MAOC and, therefore, we will include a paragraph in the discussion speculating about the possible implication of the presence of POC in the sample and how this affects the saturation estimate. We thank the reviewer for suggesting two interesting papers. We tried to apply the suggested relation presented by Feng et al. which refers to cultivated soils and gave a slightly higher saturation level. Anyway, as pointed out by the Reviewer the soils are far from saturation even including POC and this will be one main point for subsequent discussion.
- Interpretation of 4 per mille objective based on last SOC sampling only, when all treatments studied appear to decrease in SOC over time. The efficacy of 4 per mille and other natural climate solutions depends on increasing SOC levels, so the results of the time series, so long as they are based on sound inter-method comparison, would represent a repudiation of the 4 per mille at this site.
- We understand your concern and we partially agree. Indeed, if on the one hand, the SOC decline visible on the time series might repudiate the 4 per mille concept, on the other, the lower decrease observed under RI+PM might suggest that the practice might increase, or better, decrease less compared to the standard practice. To the best of our knowledge, the latter approach, namely using the 4 per mille for comparing treatments time by time, is the approach commonly used in literature. We wish to include both discussion points in our revised paper by including also a paragraph repudiating the 4 per mille at our site.

- Thoroughness and logic of various data interpretations. The authors find no effect of adding poultry manure on SOC, which is surprising given that exogenous C sources have previously been found to be highly effective in increasing local SOC levels. However, the rate of poultry manure dry matter addition is much lower than normally studied (1 Mg / ha annually; in meta-analysis of Kallenbach et al 2011 doi:10.1016/j.agee.2011.08.020, 5 Mg / ha is lowest dry matter addition category), which is not currently contextualized. The quantity of poultry manure was originally set to balance the C/N ratio for improving crop residue humification. For more clarity, we will include a sentence in the text. Near L200, the authors claim that the experimental site was not a SOC equilibrium, "confirmed by the low C/N ratios"; how can C/N ratios be used to infer a non-equilibrium state? The C/N ratio is used as a proxy to infer the soil C dynamics and very low values might indicate high mineralization activity with pauperisation of C stock. However, we can remove this sentence to improve the text's readability. Near L250, the authors claim how, in general, calcaric, silty soils are inert to management practices because they are compacted and sometimes anoxic, but how are any of these characteristics grounds for a soil being unresponsive to management practices? Those characteristics are not directly responsible for soil inertia, we will clarify this sentence.
- Near L255, the authors claim that SOC levels declined in the study as a result of 'agricultural intensification', but different levels of agricultural intensification were not clearly compared in the study, so how could this statement factor so prominently in their conclusions? This statement results from what is speculated in the discussion (LL196-201) "The experimental farm where the LTE is located was owned by the University by 1962. Before the start of the experiment, the field was conducted following the intensification of agricultural practices, typical of the first part of the 60s" [...] "Therefore, the general SOC decline might not represent the effect of a land-use change but rather a movement from a threshold level to another inside a cultivated agroecosystem, in particular a shift from a low (e.g., shallow non-inversion tillage) to more intensive agriculture (e.g., moldboard ploughing)."

  Writing style. The manuscript writing, in terms of structure and style does not yet meet a high standard of quality. Paragraph structure is not consistently used, as some sentences are presented outside of paragraphs (last sentence of introduction; second sentence of section 2.2; last sentence of section 3.1). The meaning of 'sensibly' as a modifier, repeatedly used, is unclear. Parts of the manuscript are written in an informal or casual style that should be revised in order to be suitable for publication in a scientific journal ("Anyway", "Nowadays").

  We thank the reviewer for pointing out this issue, we will use an international language service before the revised paper re-submission.