Reviewer 2:

The paper, entitled "A robust DayCent model calibration to assess the potential impact of integrated soil fertility management on maize yields, soil carbon stocks and greenhouse gas emissions in Kenya" emphasizes the importance of model calibration to enhance model accuracy. It utilizes a rich dataset from 4 sites in Kenya, an area that has been less represented/explored by many process-based models like DayCent, and thus, it provides a substantial amount valuable information. Furthermore, the paper centers its focuses on integrated soil fertility management (ISFM), maize yield, soil organic carbon, and greenhouse gas emission. Nevertheless, there are numerous concerns regarding the model calibration process (see Specific Comments section) and recommend a major revision to address these concerns before considering it for publication.

We thank you for your valuable feedback and will address the individual comments below.

General Comments:

- Line 106: it was not clear whether organic resources were applied once per year or once per season. Provide clarification.

This is specified in the next sentence. "Organic resources were applied only once a year, prior to planting for the long rainy season in January or February." However, we now also added it to the sentence you refer to.

- Section 2.3.3: Provide more detailed information on historical cropping and specify the simulation periods for reproducibility, preferably in a table format. Additionally, include information of the optimal duration of cropping systems following the transition from native condition to achieve the initial SOC levels. It would be helpful to provide a figure showing the time series of SOC stocks for the entire simulation including native condition and historic cropping systems for each site.

In response to the feedback you and reviewer 1 on the model initialization, we will now completely eliminate the spin-up and historical runs, instead relying on measured g MAOC g^{-1} SOC as a proxy for the SOC in the passive pool. Thus, this whole section will be overhauled and the table will not be necessary. (see details in comment to reviewer 1 and your first "specific comment", below)

- In Section 2.5, provide the equation for the likelihood function used in the Bayesian calibration. Additionally, clarify whether the same likelihood function was employed for the GSA, and mention this in the text.

We will provide the likelihood function for the BC. For the GSA, we did not use a likelihood function, it was based on the simulated output. This is specified in the last sentence of Section 2.4 and we overhauled the sentence, to make this clearer: "The parameter sensitivity was independently determined for the mean maize grain yield and aboveground biomass, averaged over all seasons at all sites, as well as for the SOC and soil total N stocks at the end of the simulation period."

- Line 292-293, provide reference(s) for the statement, “Due to the large number of observations and the mostly balanced dataset, the off-diagonal elements were set to 0”. Considering the higher autocorrelation in the time series for the modeled SOC stock, the statement may not hold true.

Based on your statement, we tested how the posterior would change if we include the covariance. It does in fact influence the results. Since, based on the reviewers’ comments, we have to rerun the calibration, we will use the likelihood function with the proper variance-covariance matrix in the revised paper.

- In Figure 7, the caption mentioned “variance (measurements)”. It is unclear whether the error bars represent variance, standard deviation, or 95% confidence interval. If variance is presented as error bars, this is unusual. Replace “variance” with “95% confidence interval” to main consistency consistent.

They are based on the measurement variance. We refined the statement to be clear. “Error bars represent 95% confidence intervals (measurements) and credibility intervals (simulations).”

- Figure 8 shows the difference relative to CT-N. It would be informative to show the relative differences in comparison to business-as-usual practices, as this would help identify and recommend management changes for better management practices.

Based on our field observations and discussions with local farmers and extension officers, the CT-N is in fact close to what smallholders do in the simulated regions in Kenya. For example, the average use of fertilizer in Kenya (which includes small- and large-scale farmers and all types of fertilizer) in the last two decades ranged between 30 and 50 kg ha^{-1}. (https://data.worldbank.org/indicator/AG.CON.FERT.ZS?locations=KE). We will look further into this issue at the national scale in a future publication, but this is beyond this article.

- In Table A1, include not only clay (%) but also sand (%) and silt (%) as required by DayCent for reproducibility.

Thank you for spotting this. We will add the sand (%), which thus will suffice DayCent input requirements (silt= 100%-sand-clay).

- In Figure A2, it is evident that measured SOC stock has been declining since the starting year. It would be helpful to discuss potential reasons for the decline and why model simulation is able to predict the decline.

This comment is likely referring to Figure 6, not A2. We have already discussed that soil erosion, which DayCent cannot simulate, could be the explanation. In our previous work we did regarding the SOC stocks at the simulated sites (https://soil.copernicus.org/articles/9/301/2023/), we discussed that the sites being relatively new in cultivation is another reason (Yet, this should be accounted for in the current manuscript, because DayCent includes the land-use history by allocating a relatively high proportion of the initial total SOC to the slow pool, and should thus be able to simulate this effect). We will look further into this after model recalibration.

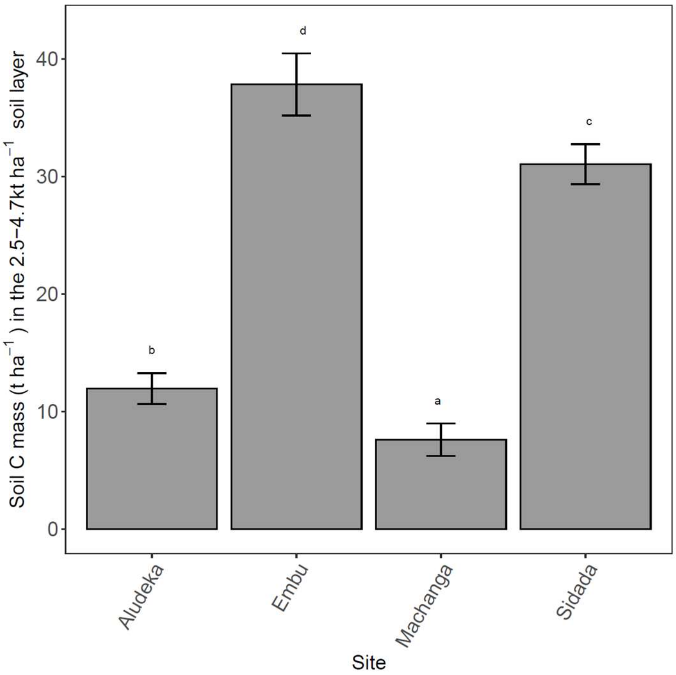
Specific Comments:

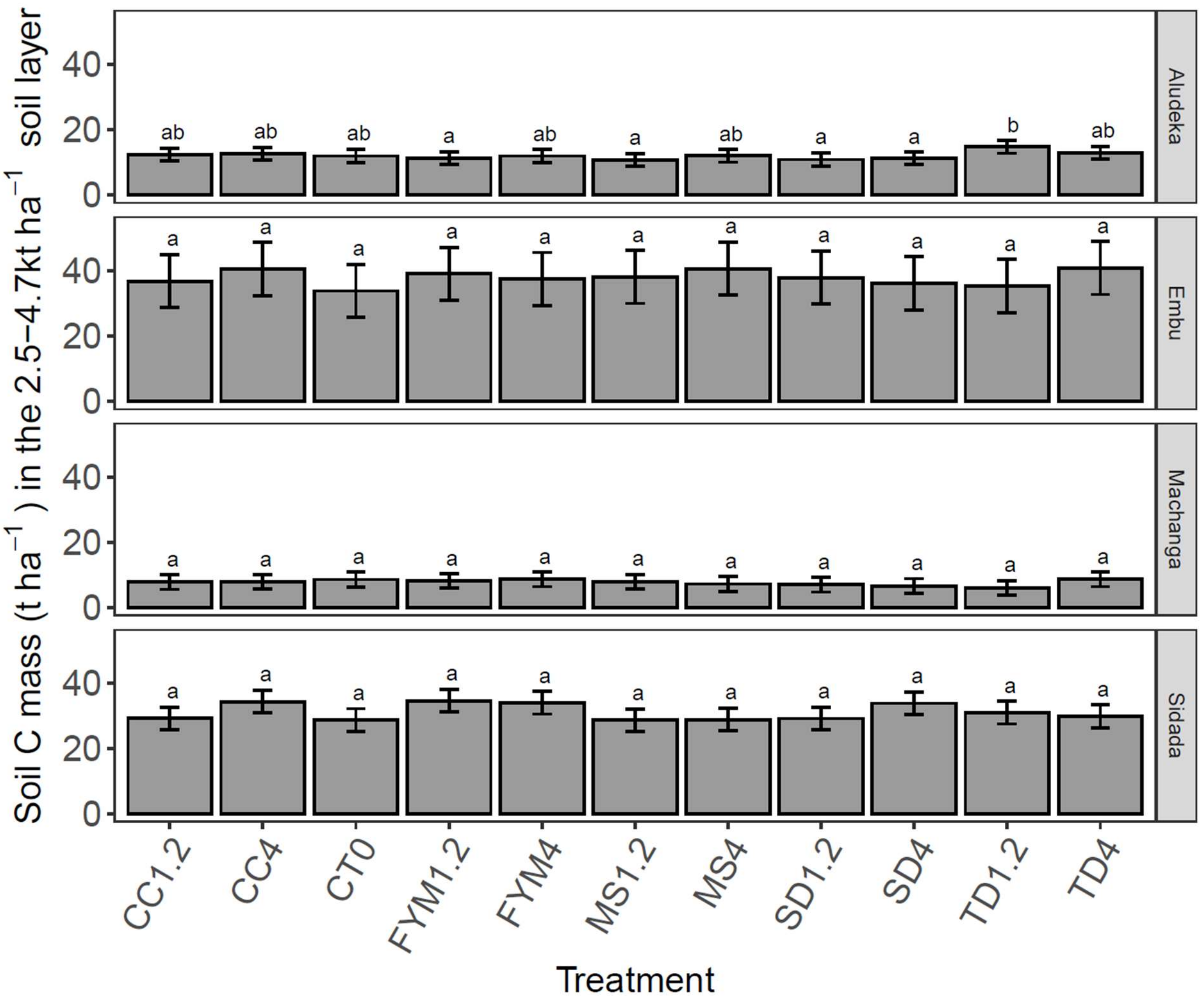
- The manuscript employs a two-step process for model predictions: Step 1 involves running the model with one set of model parameters (i.e., native condition and historical simulation) up to the beginning of experiment (i.e., initial measurement of SOC). This is done with limited adjustment to better align the model's output with measured SOC. In Step 2, a model calibration is performed, updating various parameters to a different value, with some exhibiting significant changes of several magnitude, especially the decomposition rate of slow and passive pools. Extending the model simulation with the change in parameters may disrupt the equilibrium condition and induce a drift effect, where the model attempts to reach a new equilibrium condition due to parameter changes. This makes it challenging to determine whether the changes in SOC stocks are due to alteration in management practices or change in model parameters. The potential impacts of this should be thoroughly investigated. Additionally, in line 610, the authors claims that the newly calibrated model is applicable for "upscaling the model to larger areas in Kenya" without providing practical recommendations for simulations when two sets of model parameters are available. The associated risks of such recommendations should also be examined. To mitigate potential risk, I would recommend using a model calibration procedure that results in a single set of model parameters or joint posterior distribution.

Based on this comment and others, we decided that we will eliminate the model spin-up completely – relying instead on a measured proxy of mineral-associated organic carbon (fraction of SOC that is MAOC; i.e., g MAOC g^{-1} SOC). These (unpublished) data have been measured on soil samples collected in 2021 in the framework of a master thesis in our group (the mean across treatments was 0.91, 0.88, 0.85, 0.86 g MAOC g^{-1} SOC for Aludeka, Embu, Machanga, and Sidada in 0-30 cm, respectively, with no significant treatment differences). We think this aligns with the DayCent model structure, because according to the DayCent manual, particulate organic carbon (POC) and MAOC are related (though not fully equivalent) to the slow pool and the passive pool, respectively. We will thus utilize the fraction of SOC that is MAOC to initialize the passive SOC pool of the model, while keeping the active pool at the DayCent recommended mean 3% initially, and the slow pool as the rest. It is stated in the DayCent manual that the slow pool is larger than the measured POC fraction. Consequently, the passive pools must be smaller than the g MAOC g^{-1} SOC fraction. Furthermore, we have only data from 2021, when the trials were already 19 and 16 years old. To account for these two points when using the g MAOC g^{-1} SOC fraction as a proxy to initialize the passive pool, we will add two new parameters to the Bayesian calibration: 1) an intercept and 2) a slope for time since experiment start. The intercept accounts for the fact that the passive pool is smaller than the MAOC fraction, the slope for the fact that SOC has been on a loosing trajectory since the start of the experiments and that passive pool is usually lost at the slowest rate. Hence the fraction of g MAOC g^{-1} SOC should have been lower at the start of the experiment. Therefore, the fractions have likely been shifted towards higher relative MAOC with time. We aim for a gaussian priors for these with a value of -0.3±0.1 for the intercept and -0.005±0.002 yr^{-1} for the slope. This would translate into am fraction of around 40 to 50 % of the total SOC in the passive pools at start, a bit higher than usually in DayCent, according to the manual, but in alignment with the rather recent conversion of the sites to agriculture.

- The manuscript utilizes initial parameter value for SOM decomposition, as reported in Gurung et al. (2020), which were suitable for SOC in the top 30 cm. However, the modeled SOC stocks were compared against measured SOC stocks up to a depth of 20 cm, thus resulting in a non-equivalent comparison. This inconsistency is evident in Figure A7, where the reported model predictions consistently show higher values than the measured SOC.
- IPCC recommends modeling SOC to a depth of 30 cm for GHG accounting and reporting. Since SOC measurements to 30 cm were available, it would be more appropriate to calibrate the model to simulate SOC to 30 cm, aligning it with the IPCC's recommendation.

You are right with these two comments. As a result, we will redo the model calibration for the 0-30 cm soil depth. However, data from the 15-30 cm soil depth was only available from an intense soil sampling campaign in 2021 (https://soil.copernicus.org/articles/9/301/2023/). After a statistical test for the 15-30 cm soil depth on that dataset (see two graphs below), we found that the equivalent soil mass (ESM) based SOC stocks in the 15-30 cm layer (2.5-4.7 t soil ha^{-1}) were not different between the treatments (with only one single exception in Aludeka). We will therefore derive the 0-30 cm SOC stocks by adding the site-specific value of the 15-30 cm ESM based SOC stocks from 2021 to the SOC stocks for 0-15 cm, which previously we scaled to 0-20 using the equation of Jobbágy and Jackson (2000).





- The manuscript employs a “leave-one-site-out” cross-validation approach; however, the analysis and results of the cross-validation were not presented. I recommend including some detail about the cross-validation process and its results in the manuscripts.

Most the results displayed (e.g., Fig 3, 4, 5, 6, A3, A4, A7, A8, A9) show the results from the “leave-one-site-out” cross-validation approach. We see however, that this was not formulated clearly enough and that the fact that we represent only the joint posterior parameters of all sites is confusing. We will therefore specify this more clearly in the text (and also display the 4 different posteriors by leaving one site out in the appendix)

Technical Corrections:

- Line 324: move the explanation “O___y the mean of the y-th type of measurement” below equation-9.

Thanks, we have done so!

- Line 335: mass unit for CO2eq/ha/yea) is missing.

Thanks, we added it.

- In the caption for Figure 7, replace “95% confidence intervals” with “95% credible intervals” for BC.

Thanks! We have adjusted this, as specified above.