Reviewer 1:

The paper describes the capability of DayCent model to simulate yield and SOC development of the different ISFM practices in SSA and its improvement after cal-val. So as presented, the paper is quite long and verbose, resulting quite hard to follow. The figures do not follow a chronological order and are often hard to interpret (see fig. A5). While authors report in M&M a wide description of parameters selection and initialization values which is appropriate and detailed, results are not very clear, often reporting average data which do not highlight the model's ability to reproduce the different selected managements. Also, the mismatch in N2O simulations make hard accounting the GWP here reported. Based on these premises, I recommend a major revision before to be acceptable for publication.

Thank you for your critical feedback. As a response to your concerns, we have conducted additional model runs, reconsidered most of the figures (e.g., displaying per site) and we reflect more critically the results with respect to the model's ability to reproduce the different selected managements. We also improved the simulations of N_2O. We think the article improved considerably with the changes we made, and hope you agree with this. See below our detailed responses to the individual comments.

Comments:

L118: …CH oxidation4. Typo. Thanks for spotting this. It was corrected.

L241-243: As authors state, DayCent needs to initialize the SOM pools to equilibrium using the typical input of biomass of the native vegetation. However, simulating native vegetation in SSA is not plausible since it is characterized by tropical evergreen forest, dry savanna and humid savanna that, with the only exception of savanna systems which was partly simulated in literature using the grass and tree layers, DayCent is not able to well simulate forest production (Gathany and Burke, 2012). Also, to my knowledge, DC was never tested over tropical environments. Authors should better explain what they used as vegetation for model spin-up.

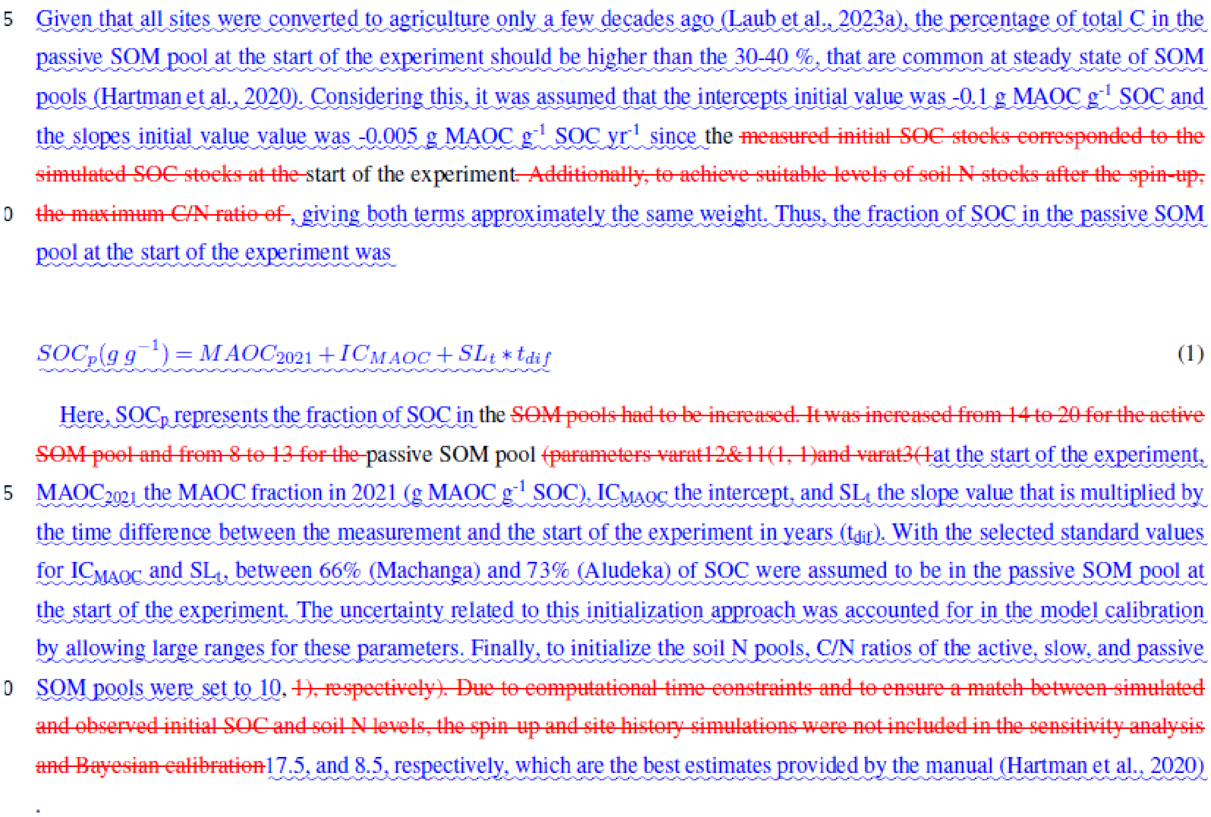
We agree that the spin-up is very uncertain for DayCent and for other similar models in general (and not just in SSA, but in general), and it was also raised as an issue by reviewer 2. Data on the history of land use is usually difficult to get in good quality (if any information is available at all), especially in SSA. This is why Mathers et al. (2023) have switched to using the spin-up and historical runs only for the distribution of total C among the different SOC pools (www.doi.org/10.1016/j.geoderma.2023.116647). However, even this comes with a lot of uncertainty regarding the real biophysical conditions and human interactions, so measured pools would in fact be best.

We therefore decided that we will eliminate the model spin-up completely – relying instead on a measured mineral-associated organic carbon pool (fraction of SOC that is MAOC; i.e., g MAOC g^{-1} SOC). See new section 2.3.3 below.

2.3.3 ~~Spin-up and site history simulation to initialize SOC and soil N contents~~Soil organic matter pools initialization based on measured data

~~As is standard practice in DayCent, the initialization of SOM pools was conducted through a~~ Instead of relying on spin-up ~~run, which was followed by a simulation of the history of the site before experiment establishment based on~~ simulation based on uncertain historical land use and management of the simulated sites, we used measured mineral associated organic carbon (MAOC) fractions as a proxy for the initialization of the ~~knowledge of site managers. The spin up has the aim to initialize the SOM pools to equilibrium using the typical input of biomass of the native vegetation type. The type of native vegetation for each site was determined from an available potential vegetation map (Kamoni et al., 2007) and confirmed by site managers as tropical evergreen forest in Embu, dry savanna in~~ passive SOM pool (Zimmermann et al., 2007). Replacing SOM initialization assumptions with measured proxies can enhance model performance (Laub et al., 2020; Wang et al., 2023) , and, more importantly, is less sensitive to user assumptions. It also aligns with the DayCent concepts on SOM; the manual (Hartman et al., 2020) denotes that particulate organic carbon (POC) and MAOC are related to the slow and the passive SOM pool, respectively. MAOC data for samples from the 0-30 cm soil layer was available from the year 2021 (specifically for the control -N, control +N and the farmyard manure -N and *Tithonia diversifolio* -N treatments at 4 t C ha^{-1} yr^{-1} at all sites). It was derived by density fractionation using sodium polytungstate solution (1.6 g cm^{-3} for Aludeka and 1.7 g cm^{-3} for the other sites). Aggregates were dispersed with ultrasonication at 400 J ml^{-1} (217 s at 200-240W), after which samples were centrifuged for 2h at 4700 rpm to separate the heavy and the light fraction, which were then separated, washed with deionised water, dried at 60°C for 24h and analyzed for weight and C content. A statistical analysis revealed the absence of treatments differences within the same site, so the site-specific MAOC values for the 0-30 cm soil depth across treatments (0.91, 0.88, 0.85, 0.86 g MAOC g^{-1} SOC for Aludeka, Embu, Machanga, and ~~humid savanna in Sidada and Aludeka. A 2000 year spin up simulation was sufficient to reach a steady state of SOM pools. Site managers had a good knowledge of the type of historical cropping systems, e. g., arable vs. grasslands, types of crop rotation (e.g., maize monoculture vs. crop rotation with legumes) , manure inputs and management, but without detailed information on the duration of these systems. Therefore, the duration of cropping systems after native vegetation was adjusted at each site so that~~ Sidada in 0-30 cm, respectively) were used to initialize the SOC in the passive SOM pool in DayCent simulations. Further, 3% of initial SOC was assigned to the active SOM pool (mean value recommended in the DayCent manual) and the remainder of SOC was assigned to the slow SOM pool.

The DayCent manual further states that, although the slow SOM pool is closely related to the POC fraction, it tends to be larger (Hartman et al., 2020). Consequently, the passive SOM pool must be smaller than the MAOC fraction. Additionally, the fractionation data was from 2021, when the experiments were already 19 and 16 years old. To address these issues, two new parameters were introduced in the simulations: 1) an intercept (IC_{MAOC}) to account for the passive SOM pool being smaller than the MAOC fraction, and 2) a slope for the time since the start of the experiment (SL_t) to account for SOM changes (mostly losses) since the start of the experiments, with the passive SOM pool typically changing at the slowest rate.



Given that all sites were converted to agriculture only a few decades ago (Laub et al., 2023a), the percentage of total C in the passive SOM pool at the start of the experiment should be higher than the 30-40 %, that are common at steady state of SOM pools (Hartman et al., 2020). Considering this, it was assumed that the intercepts initial value was -0.1 g MAOC g^{-1} SOC and the slopes initial value value was -0.005 g MAOC g^{-1} SOC yr^{-1} since the ~~measured initial SOC stocks corresponded to the simulated SOC stocks at the~~ start of the experiment~~. Additionally, to achieve suitable levels of soil N stocks after the spin-up, the maximum C/N ratio of~~ , giving both terms approximately the same weight. Thus, the fraction of SOC in the passive SOM pool at the start of the experiment was

$$SOC_p(g\,g^{-1}) = MAOC_{2021} + IC_{MAOC} + SL_t * t_{dif} \quad (1)$$

Here, SOC_p represents the fraction of SOC in the ~~SOM pools had to be increased. It was increased from 14 to 20 for the active SOM pool and from 8 to 13 for the~~ passive SOM pool ~~(parameters varat12&11(1, 1)and varat3(1~~at the start of the experiment, $MAOC_{2021}$ the MAOC fraction in 2021 (g MAOC g^{-1} SOC), IC_{MAOC} the intercept, and SL_t the slope value that is multiplied by the time difference between the measurement and the start of the experiment in years (t_{dif}). With the selected standard values for IC_{MAOC} and SL_t, between 66% (Machanga) and 73% (Aludeka) of SOC were assumed to be in the passive SOM pool at the start of the experiment. The uncertainty related to this initialization approach was accounted for in the model calibration by allowing large ranges for these parameters. Finally, to initialize the soil N pools, C/N ratios of the active, slow, and passive SOM pools were set to 10, ~~1), respectively). Due to computational time constraints and to ensure a match between simulated and observed initial SOC and soil N levels, the spin-up and site history simulations were not included in the sensitivity analysis and Bayesian calibration~~17.5, and 8.5, respectively, which are the best estimates provided by the manual (Hartman et al., 2020)

.

L335: Authors should consider replacing the term GWP with GHG balance. Despite the likely low effect of CH4, the model is not able to predict CH4 emissions, that therefore they cannot be considered in the whole balance. In this context, would be better to define the GWP as GHG balance since, in any case, the contribution of CH4 cannot be measured neither excluded.

Thanks. We have adjusted the name to GHG balance.

L338: Figure 1 is included in M&M, please move below in Results.

Thanks, we moved it.

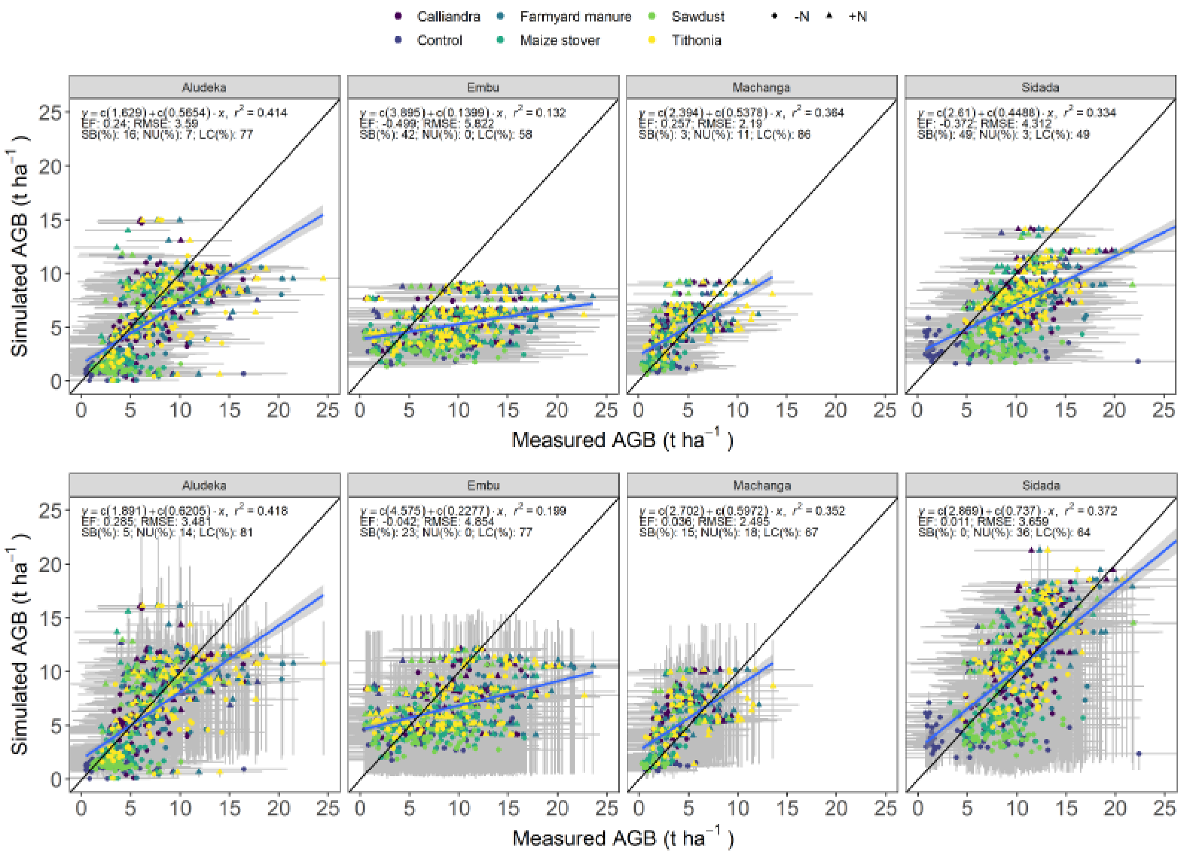
L390-393: Authors can remove this part since calibration is widely recognized to improve model performances.

We gave this suggestion some thought but in the end concluded that it is important to keep it because the model performance improvement is from leave-one-site-out cross-validation. Hence, the performance at each of the four sites was improved despite the calibration being done with only the three other sites. This is notable and indicates that the improvement of DayCent parameters suited the tropical conditions and was not an overfitting for each individual site. However, we acknowledge that we did not clearly specify that most of the results are from the leave-one-site-out cross-validation (e.g., Figures 4 to 7 are all from this leave-one-site-out cross-validation, despite combining all sites in one graph). We made this clearer now in the revised manuscript.

L394: ….and for aboveground biomass for all sites except Machanga. You mean Aludeka?

This is actually correct, it is only that we had not displayed it for AGB. (see below). As specified above, we will use the graphs per site in the new version of the article.

Uncalibrated model results on top, calibrated ones at the bottom



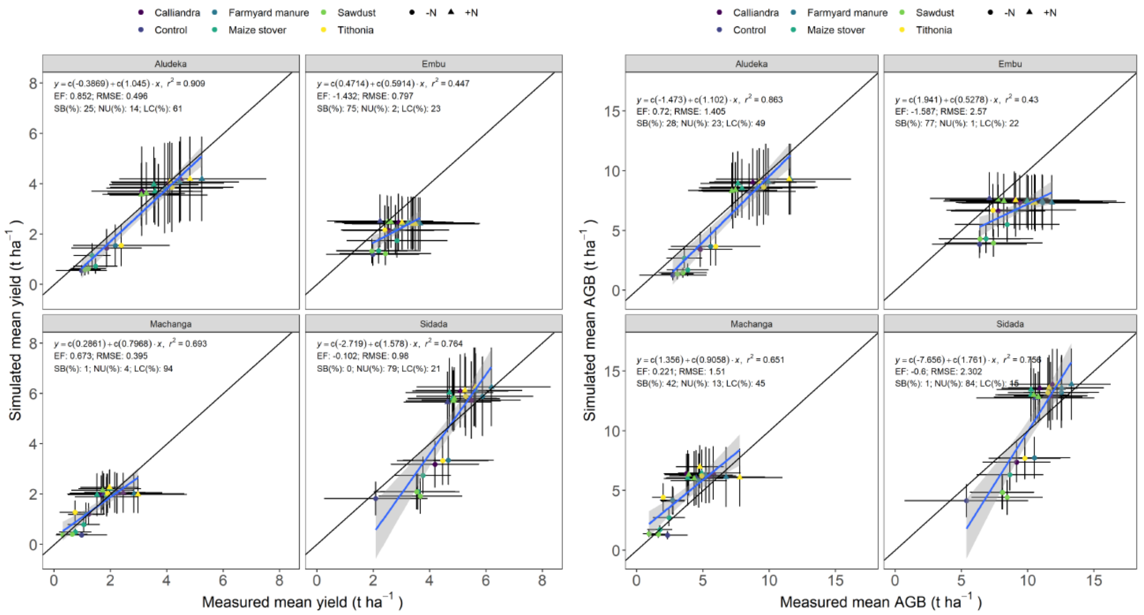
L399-401: please, when cited into the main text, report the supplementary figures in chronological order (why A9 before A4, etc…?). Also, why fig.4-5-6 in paragraph 3.5? It's quite hard to follow this flow….

Thank you for this comment because this should indeed be in order. Hence, we put special attention on the chronological order of tables and figures during revision.

Major weaknesses:

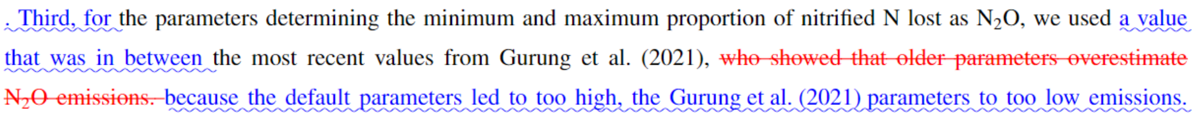
a) In Fig. 3 authors reported all together sites and management for comparing not vs calibrated model. To my opinion, this representation of model calibration is misleading. Firstly, looking at the performances for each site (Fig. A9), model calibration only little improve the model performances found using default values, with statistics confirming the improvement is quite low and lower for each site compared to when assessed overall. This confirm that averaging all sites make unclear to evaluate the model performances under different conditions. Also, it is not clear the ability of the model to reproduce different type of management after calibration process (Fig. A5 is poorly readable, and statistics should be reported. From a visual analysis, variability seem not well simulated). So, from the whole study, does not clearly emerge how the model is able to reproduce yield and AGB for each ISFM at each site. This do not allow to discuss why model does or does not work at each site and for each management, which could be the limitations and weaknesses, which should be the best practice to use and its response at each site. Averaging all yield data does not clarify the efficiency of the model to be suitable as tool to assess the potential of specific ISFM management practices (as stated by authors in introduction) to cope with food insecurity or further issues. Authors should revise all this part to provide a more accurate response to what they stated in the introduction.

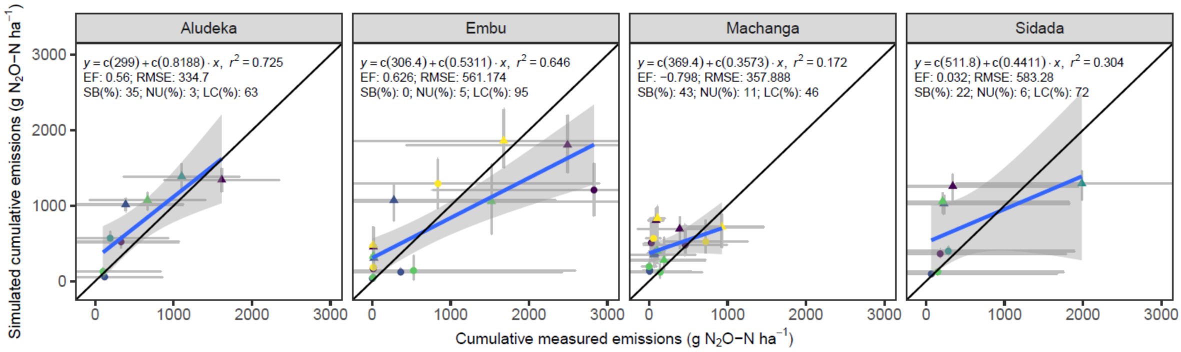
Based on your comment, we have put the site specific cross-evaluation results into the main text and report across site model statistics only in the figure captions. We now also describe in more detail, which ISFM techniques are reasonably represented by the model vs which ones are not, and for which sites and conditions the model is performing the least. We additionally added evaluation criteria to the new Fig. A4, showing that site means have a lower RSME than yearly simulated yields. Fig. A5 from the last version was removed, because it did not add any relevant information that was not presented in Figs. 3, 4, and A4, combined.



b) The GWP discussion is another major point of weakness. Results clearly showed as N2O is not well simulated neither at daily scale (Fig. A10) nor as cumulated (Fig. 7). Despite in discussions authors state that simulated N2O emissions were generally reasonably well predicted with this current DayCent calibration, looking at Fig. 7 emerged as at Aludeka and Embu the measured N emissions were more than double than those simulated. This clearly affect GWP analysis, especially considering the role of N in GWP analysis, thus making these results very uncertain. Authors should exclude GWP analysis from this study or should much better calibrate the N response to better fit with observations, otherwise GWP discussion risk to be highly speculative due to low level of confidence in N emission outcomes.

Thanks for this suggestion. After the recalibration in the revised paper (and when choosing more suitable N_2O model parameters), we could improve the cumulative N_2O predictions, removing most of the systematic model underprediction of N_2O emissions.


. Third, for the parameters determining the minimum and maximum proportion of nitrified N lost as N_2O, we used a value that was in between the most recent values from Gurung et al. (2021), ~~who showed that older parameters overestimate N_2O emissions.~~ because the default parameters led to too high, the Gurung et al. (2021) parameters to too low emissions.



We also adjusted the text as follows:

that are poorly represented in the tropics (Van Looy et al., 2017). However, the fact that cumulative N_2O emissions were better captured than daily emissions~~and~~ , that there was no systematic under- or over-prediction of cumulative N_2O emissions, ~~does suggest~~ and that simulated N_2O emissions were ~~generally reasonably well predicted~~ in the uncertainty range of measured N_2O emissions, does not provide any strong evidence against the suitability of DayCent to represent N_2O emissions with this current ~~DayCent calibration. This is important for the predictions of the GWP.~~ calibration. Because the simulation of SOC change ~~showed low bias, we can conclude that this part of the GWP is well represented. Depending on the site and treatment,~~ and cumulative N_2O emissions showed only limited bias, the GWP seem to be at least a reasonable first estimate. The contributions of N_2O emissions ~~contributed between 80~~to GWP of up to 100% (Aludeka) and ~~up to 20~~between 10 to 50% of the GWP (other sites; Fig. 9)~~. However, the~~ , are, however subject to high uncertainty, as already evident from the measurements. The larger confidence intervals of the measured compared to the simulated cumulative N_2O emissions suggest that the DayCent model cannot fully represent the variability. ~~Although~~ Thus, although DayCent's simulation of N_2O emissions is superior to using emission factors (dos Reis Martins et al., 2022), simulating N_2O emissions remains challenging and highly uncertain due to the complexity of the processes involved and the high temporal and spatial variability.

Reviewer 2:

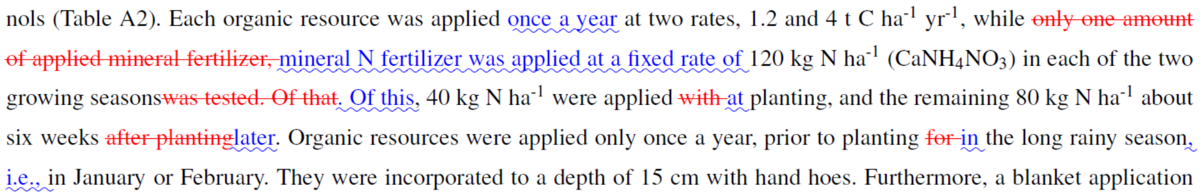
The paper, entitled "A robust DayCent model calibration to assess the potential impact of integrated soil fertility management on maize yields, soil carbon stocks and greenhouse gas emissions in Kenya" emphasizes the importance of model calibration to enhance model accuracy. It utilizes a rich dataset 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 this to the sentence you refer to.



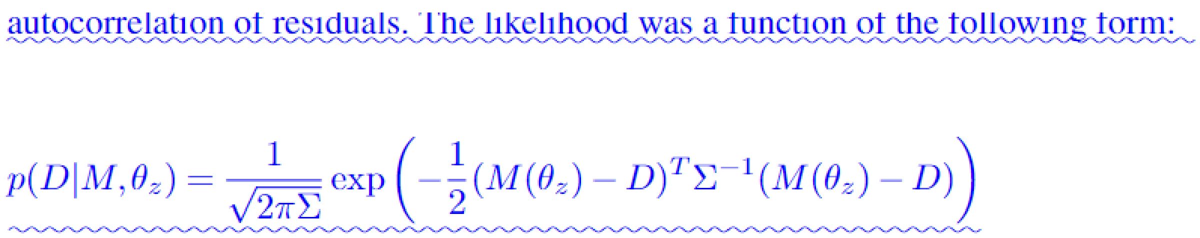
nols (Table A2). Each organic resource was applied once a year at two rates, 1.2 and 4 t C ha^{-1} yr^{-1}, while ~~only one amount of applied mineral fertilizer,~~ mineral N fertilizer was applied at a fixed rate of 120 kg N ha^{-1} ($CaNH_4NO_3$) in each of the two growing seasons~~was tested. Of that~~. Of this, 40 kg N ha^{-1} were applied ~~with~~ at planting, and the remaining 80 kg N ha^{-1} about six weeks ~~after planting~~later. Organic resources were applied only once a year, prior to planting ~~for~~ in the long rainy season, i.e., in January or February. They were incorporated to a depth of 15 cm with hand hoes. Furthermore, a blanket application

- 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 from this reviewer and reviewer 1 on the model initialization, we have completely eliminated the spin-up and historical runs, instead we relied on measured mineral-associated organic carbon pool g (MAOC g^{-1} SOC) as a proxy for the SOC in the passive pool. Thus, this whole section was rewritten and the table is no longer necessary. See details in response to reviewer 1 and first "specific comment" of this reviewer 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.

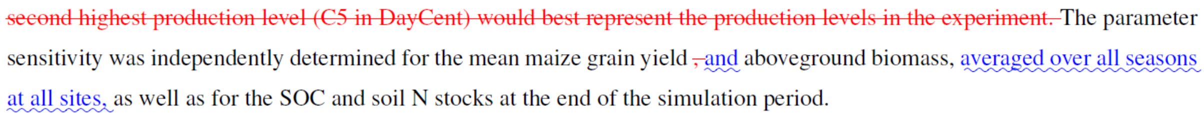
We now provide the likelihood function for the BC.



autocorrelation of residuals. The likelihood was a function of the following form:

$$p(D|M,\theta_z) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(-\frac{1}{2}(M(\theta_z) - D)^T\Sigma^{-1}(M(\theta_z) - D)\right)$$

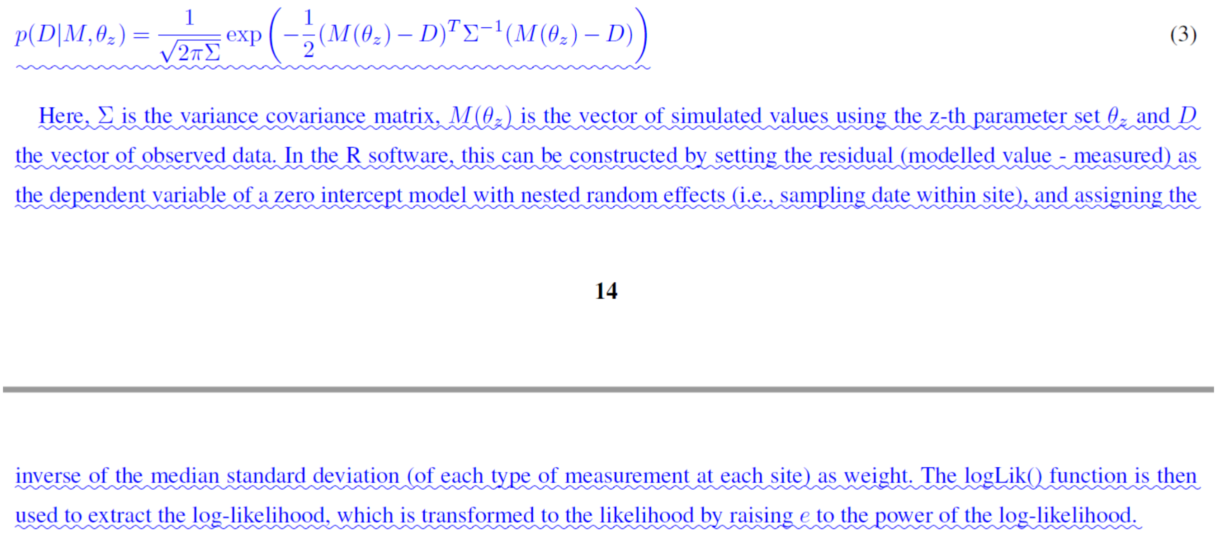
For the GSA, we did not use a likelihood function, it was based on the simulated output. This is specified in the last sentence of Section 2.4 and we rewrote the sentence, to make this clearer:



~~second highest production level (C5 in DayCent) would best represent the production levels in the experiment.~~ 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 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 included the covariance. It does in fact influence the results and we updated the likelihood function to include the covariances now.



$$p(D|M,\theta_z) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(-\frac{1}{2}(M(\theta_z) - D)^T\Sigma^{-1}(M(\theta_z) - D)\right) \quad (3)$$

Here, Σ is the variance covariance matrix, $M(\theta_z)$ is the vector of simulated values using the z-th parameter set θ_z and D the vector of observed data. In the R software, this can be constructed by setting the residual (modelled value - measured) as the dependent variable of a zero intercept model with nested random effects (i.e., sampling date within site), and assigning the

14

inverse of the median standard deviation (of each type of measurement at each site) as weight. The logLik() function is then used to extract the log-likelihood, which is transformed to the likelihood by raising e to the power of the log-likelihood.

- 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.

Thanks for highlighting this unclear description. They are based on the measurement variance. We refined the statement to ensure clarity, as follows. “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 the business-as-usual practice of smallholders in Kenya; many smallholders do not use chemical fertilizer because of being too costly (especially since the war in Ukraine) and use very minimal organic resources due to accessibility and labour constraints. Nevertheless, farmers are interested in the different ISFM treatments because they do observe soil degradation in their fields. Thus, they do want to go away from the business-as-usual scenario.

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

Thank you for bringing this to our attention. We added the sand (%), which thus meets the 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 does not simulate, could be the explanation for declining SOC. We now added the explanation "In fact, the sites were under natural vegetation (i.e. forest) or fallow up to relatively shortly before the experiment establishment. Hence upon the start of cultivation, erosion and enhanced decomposition (due to disturbance) were accelerated and have likely not yet reached a new equilibrium with C inputs from the maize. Therefore, C loss is the dominant process occurring at the sites." More details can be found in our previous work, where we discussed other reasons why SOC stocks at the simulated sites are declining (https://soil.copernicus.org/articles/9/301/2023/).

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 to eliminate the model spin-up completely – relying instead on measured mineral-associated organic carbon (i.e., fraction of SOC that is MAOC; i.e., g MAOC g^{-1} SOC). See section 2.3.3.

2.3.3 ~~Spin-up and site history simulation to initialize SOC and soil N contents~~Soil organic matter pools initialization based on measured data

~~As is standard practice in DayCent, the initialization of SOM pools was conducted through a~~ Instead of relying on spin-up ~~run, which was followed by a simulation of the history of the site before experiment establishment based on~~ simulation based on uncertain historical land use and management of the simulated sites, we used measured mineral associated organic carbon (MAOC) fractions as a proxy for the initialization of the ~~knowledge of site managers. The spin up has the aim to initialize the SOM pools to equilibrium using the typical input of biomass of the native vegetation type. The type of native vegetation for each site was determined from an available potential vegetation map (Kamoni et al., 2007) and confirmed by site managers as tropical evergreen forest in Embu, dry savanna in~~ passive SOM pool (Zimmermann et al., 2007). Replacing SOM initialization assumptions with measured proxies can enhance model performance (Laub et al., 2020; Wang et al., 2023) , and, more importantly, is less sensitive to user assumptions. It also aligns with the DayCent concepts on SOM; the manual (Hartman et al., 2020) denotes that particulate organic carbon (POC) and MAOC are related to the slow and the passive SOM pool, respectively. MAOC data for samples from the 0-30 cm soil layer was available from the year 2021 (specifically for the control -N, control +N and the farmyard manure -N and *Tithonia diversifolio* -N treatments at 4 t C ha^{-1} yr^{-1} at all sites). It was derived by density fractionation using sodium polytungstate solution (1.6 g cm^{-3} for Aludeka and 1.7 g cm^{-3} for the other sites). Aggregates were dispersed with ultrasonication at 400 J ml^{-1} (217 s at 200-240W), after which samples were centrifuged for 2h at 4700 rpm to separate the heavy and the light fraction, which were then separated, washed with deionised water, dried at 60°C for 24h and analyzed for weight and C content. A statistical analysis revealed the absence of treatments differences within the same site, so the site-specific MAOC values for the 0-30 cm soil depth across treatments (0.91, 0.88, 0.85, 0.86 g MAOC g^{-1} SOC for Aludeka, Embu, Machanga, and ~~humid savanna in Sidada and Aludeka. A 2000 year spin up simulation was sufficient to reach a steady state of SOM pools. Site managers had a good knowledge of the type of historical cropping systems, e. g., arable vs. grasslands, types of crop rotation (e.g., maize monoculture vs. crop rotation with legumes) , manure inputs and management, but without detailed information on the duration of these systems. Therefore, the duration of cropping systems after native vegetation was adjusted at each site so that~~ Sidada in 0-30 cm, respectively) were used to initialize the SOC in the passive SOM pool in DayCent simulations. Further, 3% of initial SOC was assigned to the active SOM pool (mean value recommended in the DayCent manual) and the remainder of SOC was assigned to the slow SOM pool.

The DayCent manual further states that, although the slow SOM pool is closely related to the POC fraction, it tends to be larger (Hartman et al., 2020). Consequently, the passive SOM pool must be smaller than the MAOC fraction. Additionally, the fractionation data was from 2021, when the experiments were already 19 and 16 years old. To address these issues, two new parameters were introduced in the simulations: 1) an intercept (IC_{MAOC}) to account for the passive SOM pool being smaller than the MAOC fraction, and 2) a slope for the time since the start of the experiment (SL_t) to account for SOM changes (mostly losses) since the start of the experiments, with the passive SOM pool typically changing at the slowest rate.

Given that all sites were converted to agriculture only a few decades ago (Laub et al., 2023a), the percentage of total C in the passive SOM pool at the start of the experiment should be higher than the 30-40 %, that are common at steady state of SOM pools (Hartman et al., 2020). Considering this, it was assumed that the intercepts initial value was -0.1 g MAOC g^{-1} SOC and the slopes initial value value was -0.005 g MAOC g^{-1} SOC yr^{-1} since the ~~measured initial SOC stocks corresponded to the simulated SOC stocks at the~~ start of the experiment~~. Additionally, to achieve suitable levels of soil N stocks after the spin-up, the maximum C/N ratio of~~ , giving both terms approximately the same weight. Thus, the fraction of SOC in the passive SOM pool at the start of the experiment was

$$SOC_p(g\,g^{-1}) = MAOC_{2021} + IC_{MAOC} + SL_t * t_{dif} \tag{1}$$

Here, SOC_p represents the fraction of SOC in the ~~SOM pools had to be increased. It was increased from 14 to 20 for the active SOM pool and from 8 to 13 for the~~ passive SOM pool ~~(parameters varat12&11(1, 1)and varat3(1~~at the start of the experiment, $MAOC_{2021}$ the MAOC fraction in 2021 (g MAOC g^{-1} SOC), IC_{MAOC} the intercept, and SL_t the slope value that is multiplied by the time difference between the measurement and the start of the experiment in years (t_{dif}). With the selected standard values for IC_{MAOC} and SL_t, between 66% (Machanga) and 73% (Aludeka) of SOC were assumed to be in the passive SOM pool at the start of the experiment. The uncertainty related to this initialization approach was accounted for in the model calibration by allowing large ranges for these parameters. Finally, to initialize the soil N pools, C/N ratios of the active, slow, and passive SOM pools were set to 10, ~~1), respectively). Due to computational time constraints and to ensure a match between simulated and observed initial SOC and soil N levels, the spin-up and site history simulations were not included in the sensitivity analysis and Bayesian calibration~~17.5, and 8.5, respectively, which are the best estimates provided by the manual (Hartman et al., 2020)

.

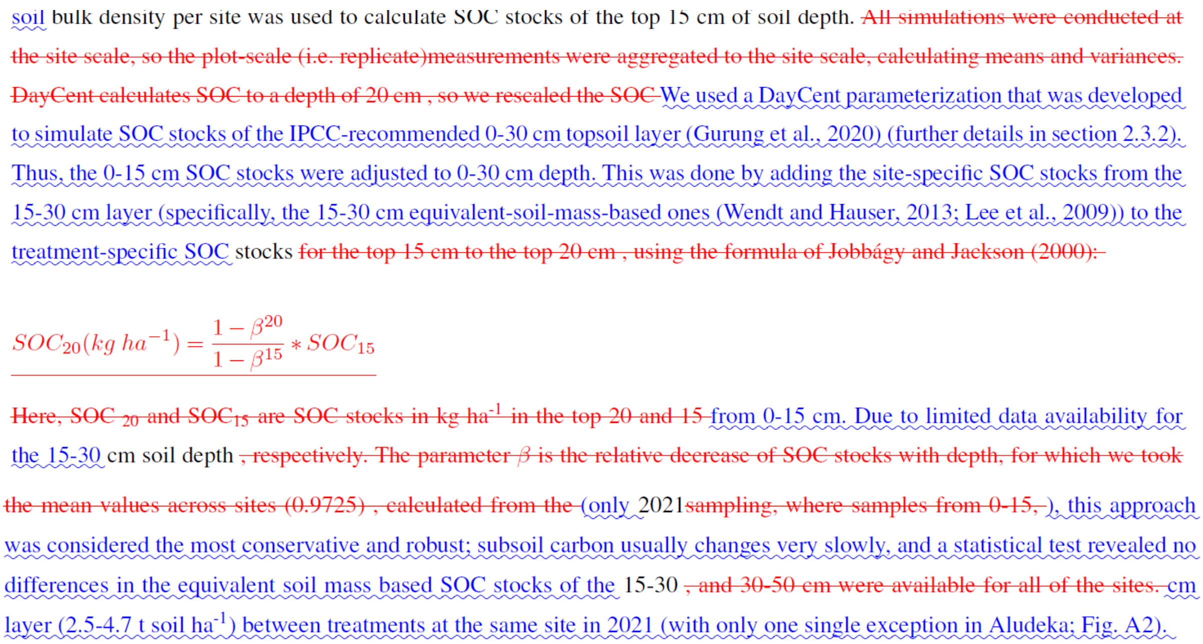
We further added the requested recommendations for potential upscaling exercises, that in the best case the full posterior parameter set should be used to derive the uncertainties of estimates, and that similar errors can be expected.

Because our calibration shows a good ~~fit and is free from serious bias even for~~ model fit with observed mean yields and changes in SOC stocks ~~(that is, 88% of the errors are LCand not systematic; EF is 0.55~~across sites, with no overall major bias (positive EF and errors mostly consisting of LC), the ~~calibrated DayCent model can be used in a robust manner to estimate the change of SOC stocks under different ISFM management in Kenya. Furthermore, because the model evaluation and calibration~~ parameter set, especially the full posterior, appears suitable for upscaling of model simulations. However, on should keep in mind that the season-to-season yield variability is captured less accurately than the mean yields (lower RMSE) and that changes in SOC are better represented at sites with clay-rich soils than those with clay-poor soils. Because the model calibration and evaluation were performed at ~~different sites , it seems reasonable to use DayCentfor other~~ sites with diverse characteristics, it is reasonable to assume that DayCent, when applied to sites with similar climate and soil conditions, ~~even beyond Kenya~~will provide satisfactory results with similar model uncertainties and errors. In that respect, while the leave-one-site-out cross-validation ~~is using the data most efficiently: for evaluation, we leave one site out, but for~~ made efficient use of data for model evaluation, further model upscaling ~~exercises, ideally~~ should apply the full posterior model parameter set including all sites (Fig. 2) should be used. ~~A computationally less expensive alternative is to~~ In that case, a computationally inexpensive exercise would use only the single ~~parameter set with the highest likelihood~~ best parameter set (Table 1), while the full posterior parameter set should be used to get estimates of the posterior credibility intervals for changes in SOC stocks.

- 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.

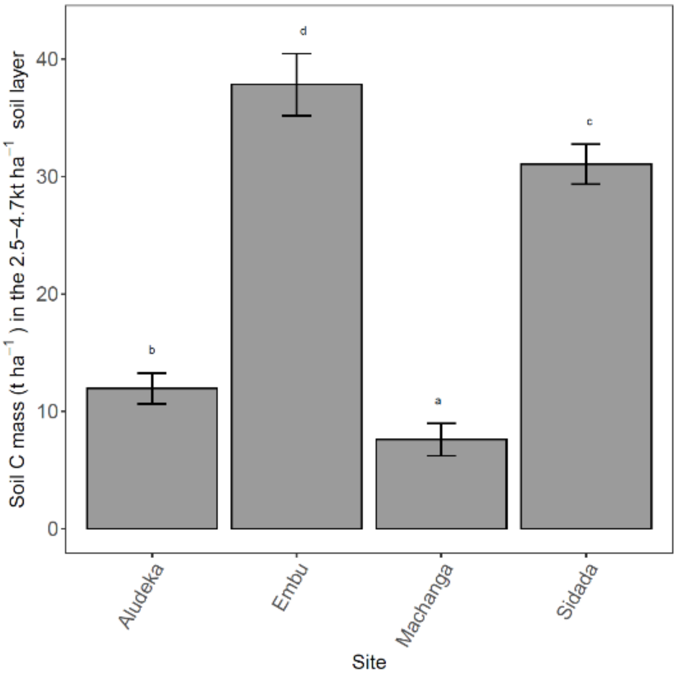
We agree with these two important comments. As a result, we have redone the model calibration, now using data for the 0-30 cm soil depth. See section 2.3 pasted below with associated figures:

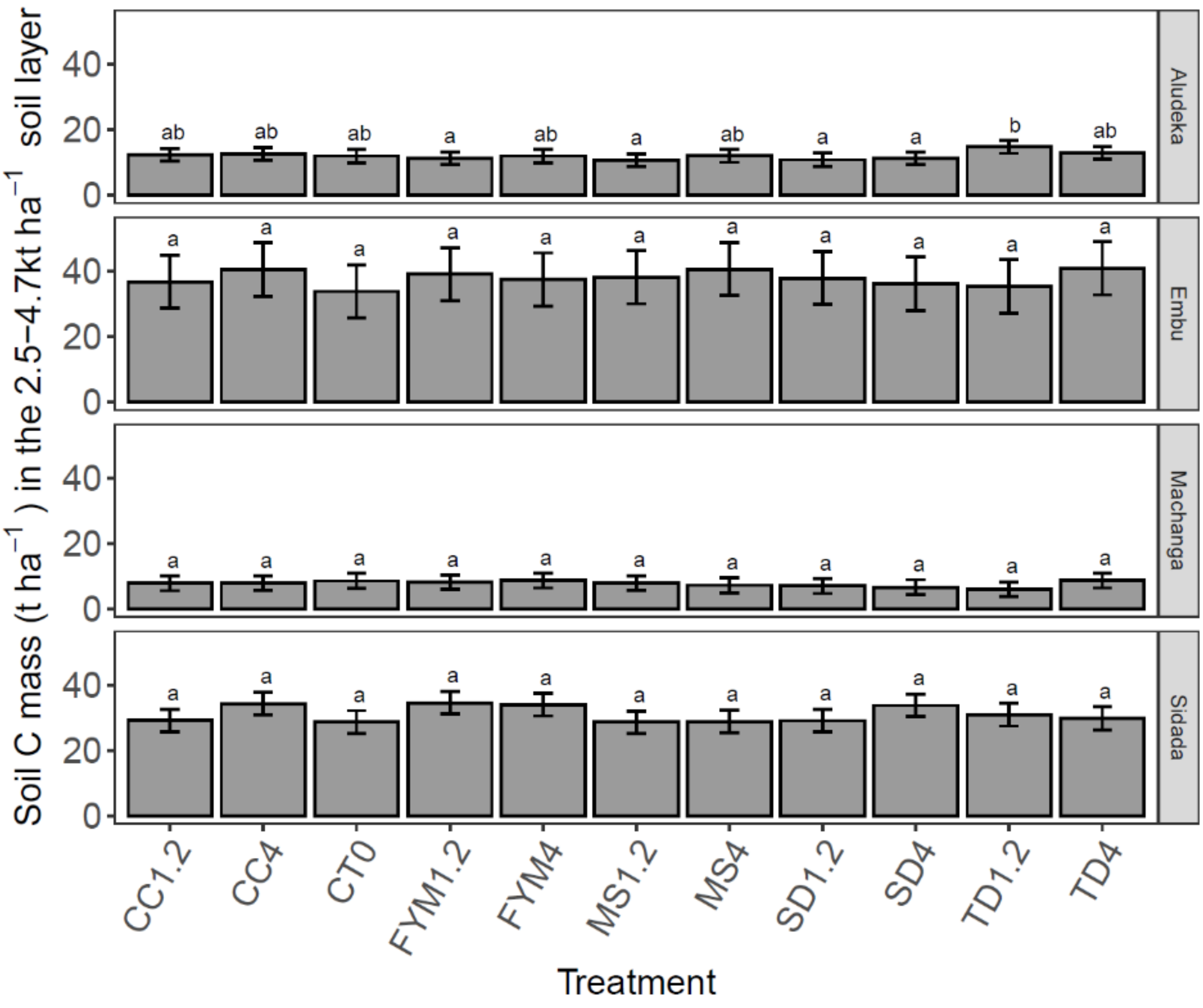


soil bulk density per site was used to calculate SOC stocks of the top 15 cm of soil depth. ~~All simulations were conducted at the site scale, so the plot-scale (i.e. replicate)measurements were aggregated to the site scale, calculating means and variances. DayCent calculates SOC to a depth of 20 cm , so we rescaled the SOC~~ We used a DayCent parameterization that was developed to simulate SOC stocks of the IPCC-recommended 0-30 cm topsoil layer (Gurung et al., 2020) (further details in section 2.3.2). Thus, the 0-15 cm SOC stocks were adjusted to 0-30 cm depth. This was done by adding the site-specific SOC stocks from the 15-30 cm layer (specifically, the 15-30 cm equivalent-soil-mass-based ones (Wendt and Hauser, 2013; Lee et al., 2009)) to the treatment-specific SOC stocks ~~for the top 15 cm to the top 20 cm , using the formula of Jobbágy and Jackson (2000):~~

$$SOC_{20}(kg\ ha^{-1}) = \frac{1-\beta^{20}}{1-\beta^{15}} * SOC_{15}$$

~~Here, SOC $_{20}$ and SOC$_{15}$ are SOC stocks in kg ha^{-1} in the top 20 and 15~~ from 0-15 cm. Due to limited data availability for the 15-30 cm soil depth ~~, respectively. The parameter β is the relative decrease of SOC stocks with depth, for which we took the mean values across sites (0.9725) , calculated from the~~ (only 2021~~sampling, where samples from 0-15,~~), this approach was considered the most conservative and robust; subsoil carbon usually changes very slowly, and a statistical test revealed no differences in the equivalent soil mass based SOC stocks of the 15-30 ~~, and 30-50 cm were available for all of the sites.~~ cm layer (2.5-4.7 t soil ha^{-1}) between treatments at the same site in 2021 (with only one single exception in Aludeka; Fig. A2).





- 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.

We acknowledge that we had not stated clearly enough that all plots of simulated compared to measured data were effectively from the “leave-one-site-out” cross-validation. This has been clarified, and we also show the different posteriors from leaving out each site. Based on the comments from reviewer 1, we also have moved the evaluation graphs by-site into the main text.

2.3 Data used for the DayCent model ~~evaluation/~~calibration and evaluation

~~In a process that was repeated four times, a large data set was used based on~~

To provide an overall assessment of the performance of DayCent for its use in Kenya a leave-one-site-out cross-validation approach was applied. Specifically, this involved using a data sub-set from three of the four sites for ~~the model calibrationand the validation was performed based on the~~ model calibration, with validation performed using the data from the fourth site. ~~The plot-scale yield of maize grain~~ This process was repeated four times, every time with another site serving as the validation site. Different data, were used for this: Maize grain yield and the aboveground biomass, both on a dry matter basis~~(t ha^{-1})~~,

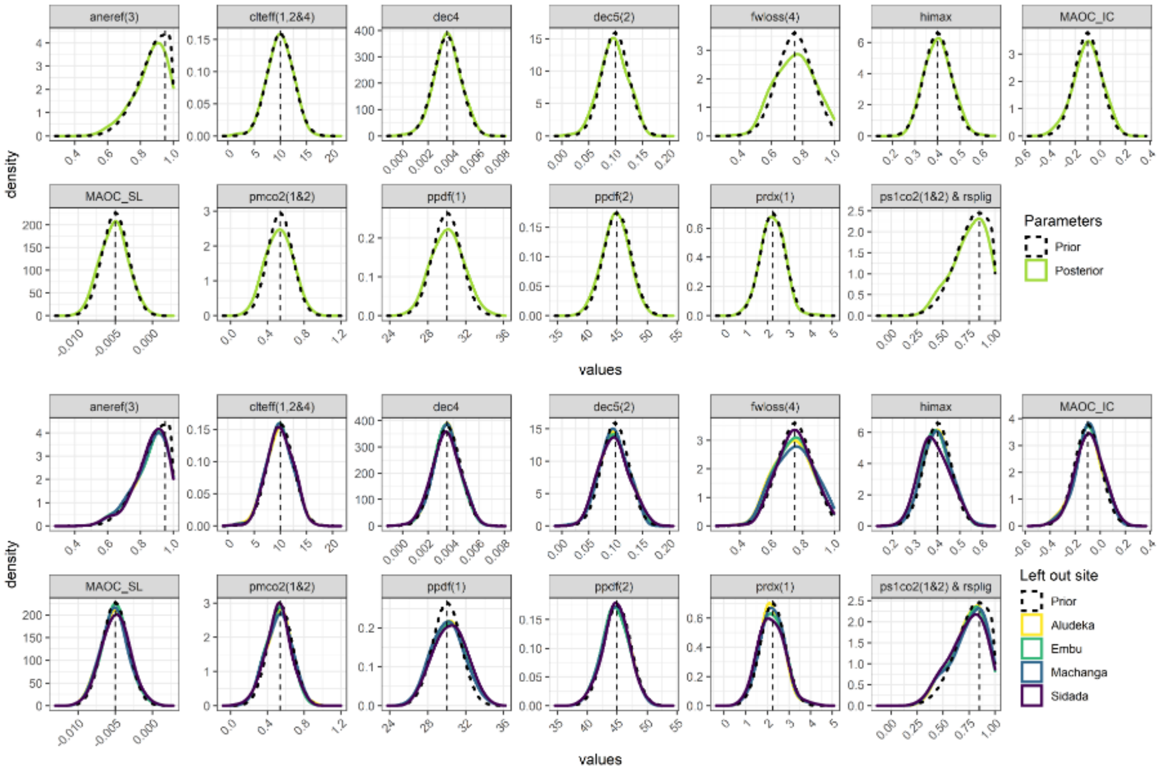


Figure 2. Prior compared to the posterior model parameter distribution resulting from the uncertainty-based Bayesian model calibration of DayCent using data from all sites combined (top) and the leave-one-site-out cross-validation (bottom). Dashed vertical lines represent the values of the ~~default~~ initially selected parameter set. The posterior distributions are based on all four study sites combined. For the description of the parameters see Table 1.

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.

Community Comment 1:

General comments: Very nice paper. Generally well written. The M&M in particular are very thorough. I have not done much modeling, but I found that the M&M did a good job of explaining the model parameters and their calibration along with how sensitive they were. Apart from a bunch of small issues (see below), I found that the discussion around objective iii. was lacking a bit. What I was really looking forward to was more discussion around the trade-offs between yield and SOM / increases along with the global warming potential of the different ISFM treatments.

Thanks for this positive feedback. We refined the discussion part on objective iii, after a model recalibration, based on the comments of reviewer 1 and 2. However, since there was still quite some uncertainty around simulated N_2O emissions, we focused this discussion section on this uncertainty and its potential sources.

Specific comments:

Lines 85-90: A map with the site locations would be helpful here as well.

We agree that a map would be helpful but decided against having one, since the manuscript already contains a lot of figures. However, we added a map with the locations to the Supplementary Materials.

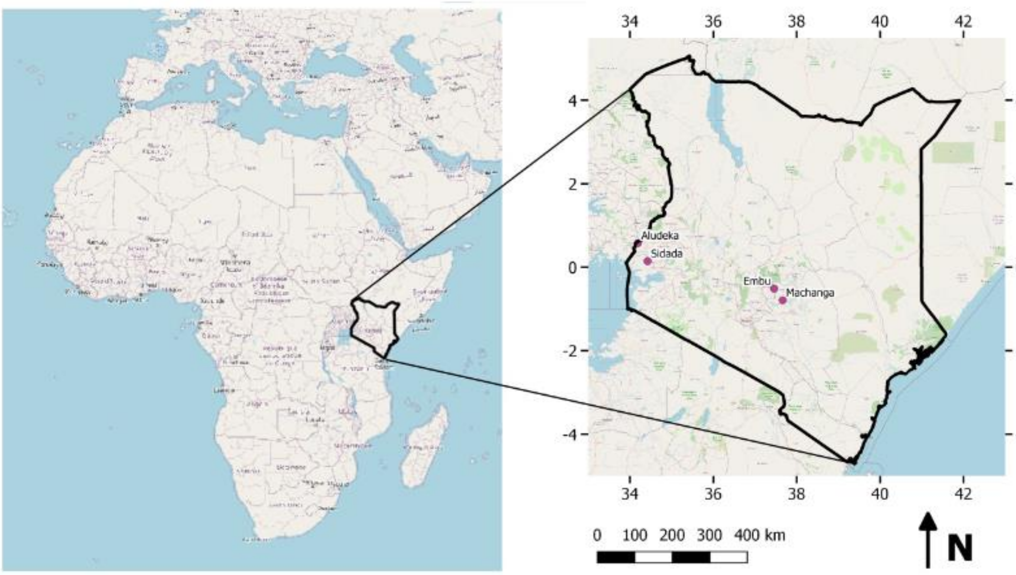


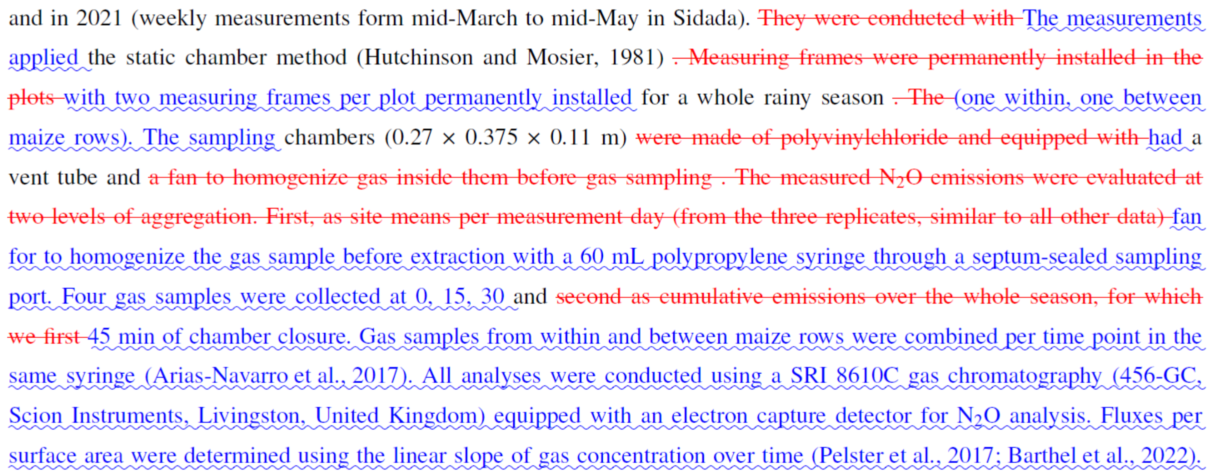
Figure A1. ~~Prior compared to posterior parameter distribution resulting from increasing~~ Map displaying the ~~ranges~~ location of the ~~uncertainty based Bayesian calibration. Dashed vertical lines represent the default parameter sets. The posterior distributions are based on all~~ four study sites~~combined~~.

Line 118 : should be “CH4 oxidation”.

Thanks, this was corrected.

Lines 150-155: How many samples per chamber? How long was the deployment time? How did you calculate the change in mixing ratios over time (linear or non-linear?), how were gas samples analyzed? (on a GC? What kind?). You need a bit more detail here.

Thanks for making us aware of this, we added the missing information to the manuscript, while at the same time trying to be brief.



and in 2021 (weekly measurements form mid-March to mid-May in Sidada). ~~They were conducted with~~ The measurements applied the static chamber method (Hutchinson and Mosier, 1981) ~~. Measuring frames were permanently installed in the plots~~ with two measuring frames per plot permanently installed for a whole rainy season ~~. The~~ (one within, one between maize rows). The sampling chambers (0.27 × 0.375 × 0.11 m) ~~were made of polyvinylchloride and equipped with~~ had a vent tube and ~~a fan to homogenize gas inside them before gas sampling . The measured N_2O emissions were evaluated at two levels of aggregation. First, as site means per measurement day (from the three replicates, similar to all other data)~~ fan for to homogenize the gas sample before extraction with a 60 mL polypropylene syringe through a septum-sealed sampling port. Four gas samples were collected at 0, 15, 30 and ~~second as cumulative emissions over the whole season, for which we first~~ 45 min of chamber closure. Gas samples from within and between maize rows were combined per time point in the same syringe (Arias-Navarro et al., 2017). All analyses were conducted using a SRI 8610C gas chromatography (456-GC, Scion Instruments, Livingston, United Kingdom) equipped with an electron capture detector for N_2O analysis. Fluxes per surface area were determined using the linear slope of gas concentration over time (Pelster et al., 2017; Barthel et al., 2022).

Line 367: Is “langley” an SI unit? I had to do an internet search to find out what it is. Would it be possible to explain what this is? Or convert to SI units?

No, but it is the unit used in DayCent. We added the explanation “(1 langley is 41 840 J m^{-2})” to the sentence.

Figure 2: shouldn’t there be some label on the X and Y axes?

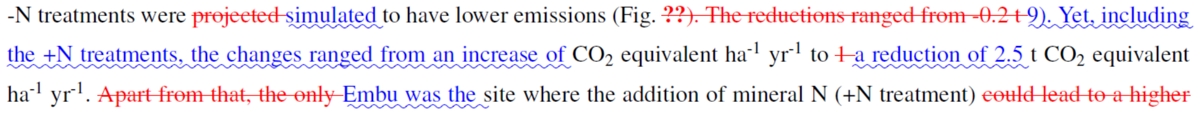
Yes, thanks! Should be "Density" and "value" as y and x axes. We added this to the new version of the manuscript.

Line 395: perhaps I don’t quite understand, but isn’t the systemic underestimation at high yields (and AGB) a “bias”?

A systematic underestimation at high yields (and AGB) is shown by the non-unity slope (<1; overestimation at low values, underestimation at high values). A bias is considered as the case of an over- or underestimation across the full range of data. Due to the large amount of data (many overlaps in the average yields), visual inspection of bias etc. are misleading and thus the SB, NU, LC assessment of Gauch (2003) are a better approach to assess whether bias exists.

Line 446: wouldn’t a negative reduction be an increase?

Yes, we reformulated, as follows:



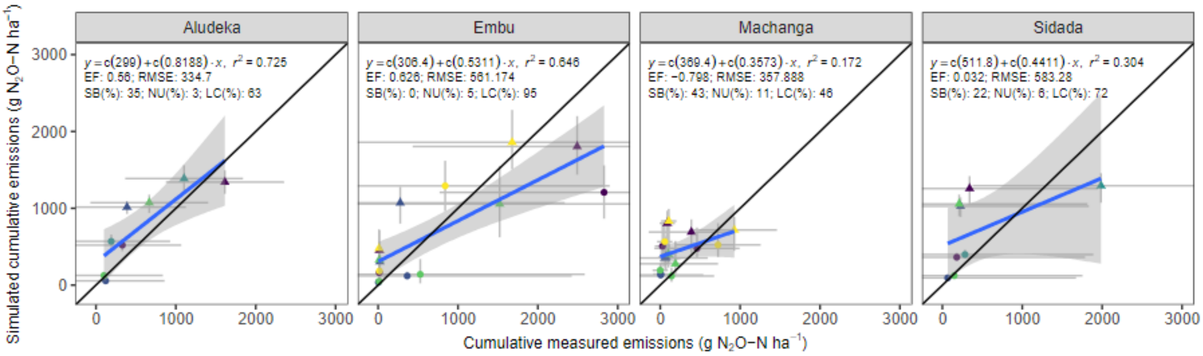
-N treatments were ~~projected~~ simulated to have lower emissions (Fig. ~~??). The reductions ranged from -0.2 t~~ 9). Yet, including the +N treatments, the changes ranged from an increase of CO_2 equivalent ha^{-1} yr^{-1} to ~~1~~ a reduction of 2.5 t CO_2 equivalent ha^{-1} yr^{-1}. ~~Apart from that, the only~~ Embu was the site where the addition of mineral N (+N treatment) ~~could lead to a higher~~

Line 447: I would say “led to” rather than “could lead to”. Since there was a reduction noted.

Thanks, we changed the text accordingly.

Figure 7: It seems that you are unable to simulate the high emission days, which could be why the cumulative simulated emissions are typically lower than the 1:1 line. Also, in the Sidada site for simulate vs measured cumulative emissions, you have one data point that has a lot of leverage. I would consider seeing how the regression line looks without that point. And maybe investigate why that point is so different from the rest of the data at that site.

Thanks for this suggestion. We have improved the simulation of N_2O in the revised version, by fitting more suitable model parameter values for N_2O production, effectively removing this bias that you and reviewer 1 mentioned (see figure below).



Lines 483-484: Mention here that DayCent overestimated SOC pre-calibration, but after the calibration the SOC concentrations (or stocks) were simulated much more accurately. This difference is clear when you look at the figure, but since the figure is in the appendix, it may not be readily apparent to the readers.

The whole section changed due to the new results, hence this part was removed.

Line 513: why do you use 2^1 and 2^3 here? Why not just say 2 and 8? Or am I missing something?

The whole section changed due to the new results, hence this part was removed.

Lines 526 to 529: Is there a reason why you switch between SOC and SOM? It seems like you are talking about the same thing.

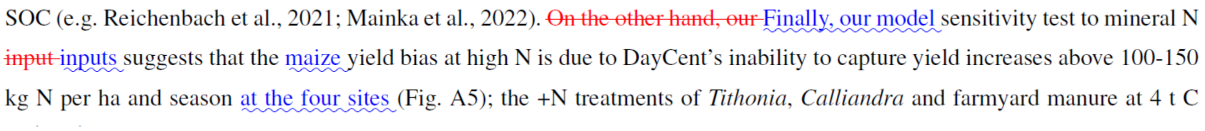
Thanks for spotting this, it should all be SOM here. SOM is the model pool, SOC is related to carbon. We went through the manuscript to correct this everywhere.

Line 534: “vary” not “very”.

Thanks, we changed this.

Line 543: Are you saying that DayCent does not capture yield increases above 100-150 kg N per ha per season in general? Or just specifically in Kenya. I have not used DayCent, but I would be very surprised if it does not capture yield increases above 150 kg N per ha in temperate regions.

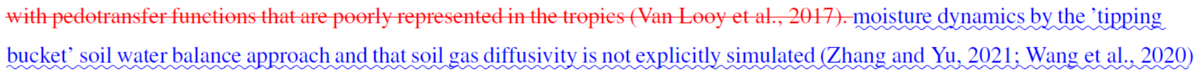
No, just in our study – we have not tested it for other sites/climates. We added “at the four sites” at the end of the sentence.



SOC (e.g. Reichenbach et al., 2021; Mainka et al., 2022). ~~On the other hand, our~~ Finally, our model sensitivity test to mineral N ~~input~~ inputs suggests that the maize yield bias at high N is due to DayCent’s inability to capture yield increases above 100-150 kg N per ha and season at the four sites (Fig. A5); the +N treatments of *Tithonia*, *Calliandra* and farmyard manure at 4 t C

Line 549-553: I wouldn't worry too much about the poor match between simulated and measured daily fluxes. I would mention though that the timing of peak fluxes is related more to soil gas diffusivity and that soil hydraulics are more just a proxy of the diffusivity.

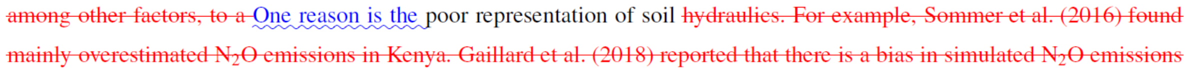
Thanks for this suggestion. We added this suggestion to the text.



~~with pedotransfer functions that are poorly represented in the tropics (Van Looy et al., 2017).~~ moisture dynamics by the 'tipping bucket' soil water balance approach and that soil gas diffusivity is not explicitly simulated (Zhang and Yu, 2021; Wang et al., 2020)

Line 553: Sommer et al. 2016 does not quite say this. What they say is that "As such, the overall model fit was exceptionally good, even though the visual impression would suggest a significant overestimation of emissions by CropSyst". If you look at the figures in their study, the simulated line up very well with the measured emissions. It is just that there are a lot of peaks in the simulated that occur between samplings.

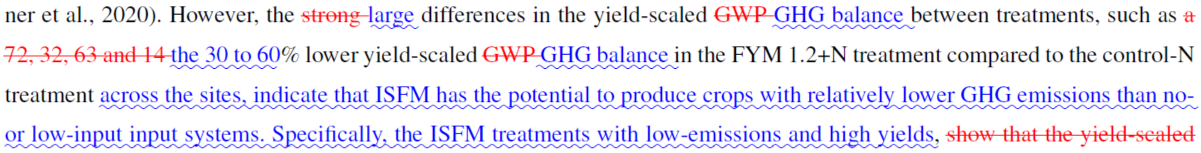
We reconsidered this part of the sentence and you are correct with regards to the text of Sommer et al (2016). We thus decided to remove the citation.



~~among other factors, to a~~ One reason is the poor representation of soil ~~hydraulics. For example, Sommer et al. (2016) found mainly overestimated N_2O emissions in Kenya. Gaillard et al. (2018) reported that there is a bias in simulated N_2O emissions~~

Line 569: I guess this is somewhat true, in that maize mono-cropping will still produce some GHG emissions. However what is the difference between the ISFM practices and the "typical" treatment (what is typical? No inputs? No N input and a small amount of FYM)? It seems like adding some inorganic N with 1.2 T C increased yields, without increasing yield scaled emissions compared with 0N 0C and compared with 0n 1.2T C. So even though it is not exactly "negative emission technology" it still seems to be an improvement.

Yes, we agree. To highlight this, we adjusted the sentence:



ner et al., 2020). However, the ~~strong~~ large differences in the yield-scaled ~~GWP~~ GHG balance between treatments, such as ~~a 72, 32, 63 and 14~~ the 30 to 60% lower yield-scaled ~~GWP~~ GHG balance in the FYM 1.2+N treatment compared to the control-N treatment across the sites, indicate that ISFM has the potential to produce crops with relatively lower GHG emissions than no- or low-input input systems. Specifically, the ISFM treatments with low-emissions and high yields, ~~show that the yield-scaled~~

Line 570: why say "positive absolute" in stead of just "positive"?

We agree and changed it to "net positive".

Lines 578-580: While I agree that N fertilizer should only be applied to responsive soils, I'm not sure that is a conclusion of the date that you have here. If you look at yields, all the sites respond to N fertilizer (either mineral or organic). It is just that they seem to respond a bit differently, particularly in the N2O emissions, to the fertilizer applications. Besides, the 0N control also has much higher yield scaled GWP in Embu and Machanga, mainly related to loss of SOC, so I don't think the higher yield scaled emissions (compared with Sidada and Aludeka) with the +N treatments indicate that these shouldn't be fertilized. In fact, the decrease in yield-scaled GWP when adding N is greater at the sites in Central Kenya than they are at Sidada, which almost contradicts what you are saying here.

We fully agree and thus removed this sentence.

Line 610: Just mention which treatment had the lowest yield-scaled emissions (the mix of FYM +N) as the preferred INMS for Kenya.

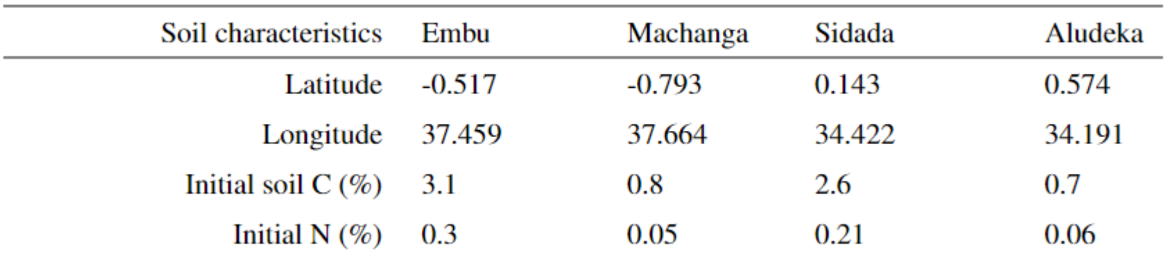
After the model recalibration, we do not think that this comes out clear enough to put it into the conclusion. We hope this is agreeable to the editor and reviewers.

Table A1: can you add the sand content as well?

Done

Figure A3: what depth are you using to calculate the stocks? You mention 15 cm depth in some locations, but you also mention that DayCent uses 20 cm depth. And, I am having a hard time seeing how the Machanga site lost so much of its C. at 20 cm depth a soil with a C content of 0.3 and a BD of 1.51 would have about 10 t C per ha. And you are saying here that it lost about 10 t per ha (or essentially all of its soil C). Is my math off (wouldn't be the first time).

We added the soil depth to the Figure (now A7). To your question of Machanga – the site had initially about 20 t C ha^{-1}. We realized that the initial soil C and N in the Table A1 were still data from the original reference profile plot description, which consisted of a single measurement per horizon at each site, conducted before the trials were established. However, at the time of trial establishments, further soil C and N measurements were done in each plot, resulting in slightly different initial C and N contents (see below). We had actually used these more accurate measures of SOC to match the SOC stocks in DayCent. We now updated Table A1 for consistency. Thanks for bringing this to our attention!



Soil characteristics	Embu	Machanga	Sidada	Aludeka
Latitude	-0.517	-0.793	0.143	0.574
Longitude	37.459	37.664	34.422	34.191
Initial soil C (%)	3.1	0.8	2.6	0.7
Initial N (%)	0.3	0.05	0.21	0.06

Figure A6: the figure caption needs to be re-done. For example, the second sentence is missing a word somewhere (perhaps "was" before "insensitive"?). And secondly, are you sure about the 50/50 split application? You were calibrating to data where the split application was 40 kg N at planting and 80 kg after ~ 6 weeks (see line 107-108; also line 165).

Thanks for bringing this to our attention. We have rewritten the caption. We are sure, about the evenly split application – this was for the technical reason that DayCent did not allow to go for higher N applications than 200kg N per one application (and we wanted to test until 400).

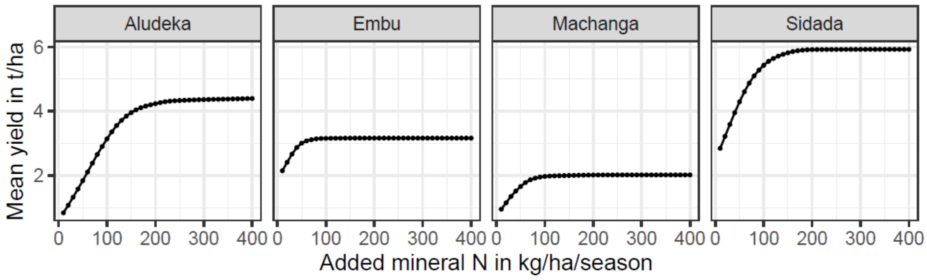
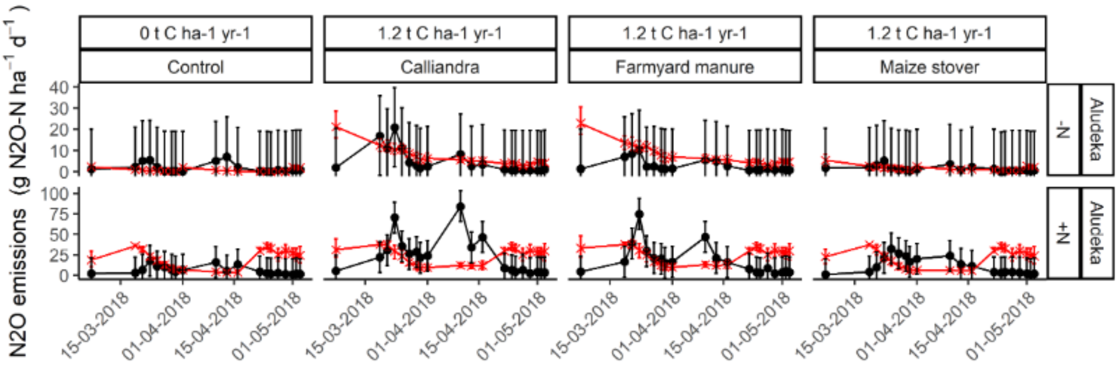
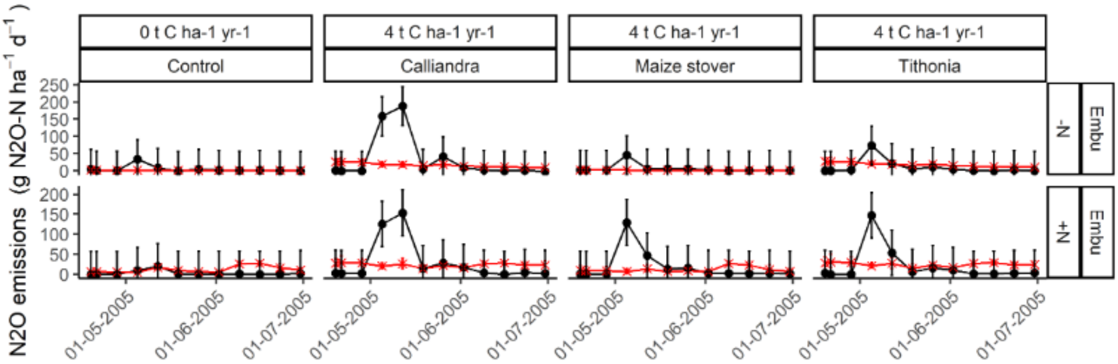


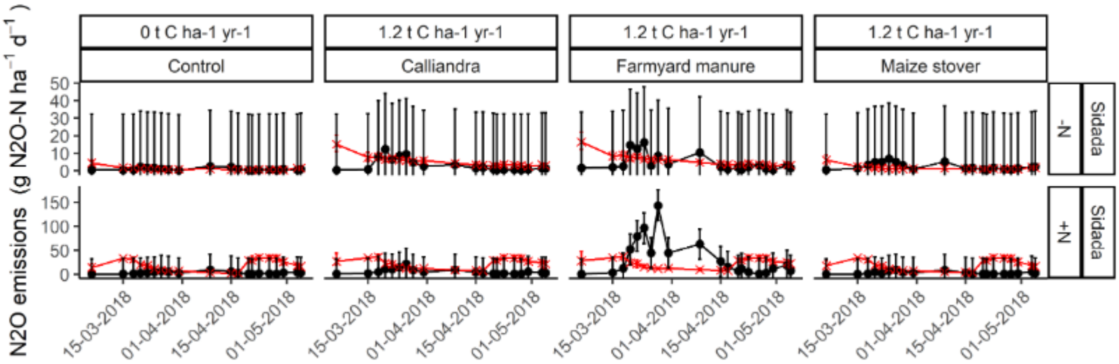
Figure A5. ~~N~~ Yield response curve ~~by site~~ of DayCent to varying levels of mineral N application (control + N treatment, without organic resources) using the calibrated DayCent parameters. ~~This was done to explain why DayCent insensitive at high N levels~~Displayed are the simulated mean yields across all simulated seasons (32 in Sidada and Aludeka, 38 seasons in Embu and Machanga). The amount of mineral N ~~was given~~ applied per season in ~~50/50~~ the simulations was evenly split ~~application, at~~ between the ~~two real~~ actual application dates of mineral N in each season at each site.

Figure A10, can you increase the font size in the figure please?

We have done so and removed the figures from Machanga, because we realized that due to erosion and runoff issues, the N_2O and SOC dynamics at this site were not very well represented by DayCent.


N2O emissions (g N2O-N ha-1 d-1)
0 t C ha-1 yr-1
Control
1.2 t C ha-1 yr-1
Calliandra
1.2 t C ha-1 yr-1
Farmyard manure
1.2 t C ha-1 yr-1
Maize stover
N-
N+
Aludeka


N2O emissions (g N2O-N ha-1 d-1)
0 t C ha-1 yr-1
Control
4 t C ha-1 yr-1
Calliandra
4 t C ha-1 yr-1
Maize stover
4 t C ha-1 yr-1
Tithonia
N-
N+
Embu


N2O emissions (g N2O-N ha-1 d-1)
0 t C ha-1 yr-1
Control
1.2 t C ha-1 yr-1
Calliandra
1.2 t C ha-1 yr-1
Farmyard manure
1.2 t C ha-1 yr-1
Maize stover
N-
N+
Sidada