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 will do further model runs, reconsider individual figures (e.g. A5) for their interpretability and present clearer results with respect to the model's ability to reproduce the different selected managements, also displaying results per site in the main text. We will shorten the text, where possible. We will also reconsider if it makes sense to exclude $N_2O$ from the GHG emissions for comparing treatments after a recalibration. See below our detailed responses to 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 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 a passive pool estimate of around 40 to 50 % of the total SOC at the start of the experiment, a bit higher than usually in DayCent, according to the manual, but in alignment with the rather recent conversion of the sites to agriculture.

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 will adhere to this suggestion. Depending on how well the $N_2O$ is predicted for the existing data after a needed recalibration, we might focus entirely on $CO_2$ and remove $N_2O$ as well.

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

Thanks, we moved it.

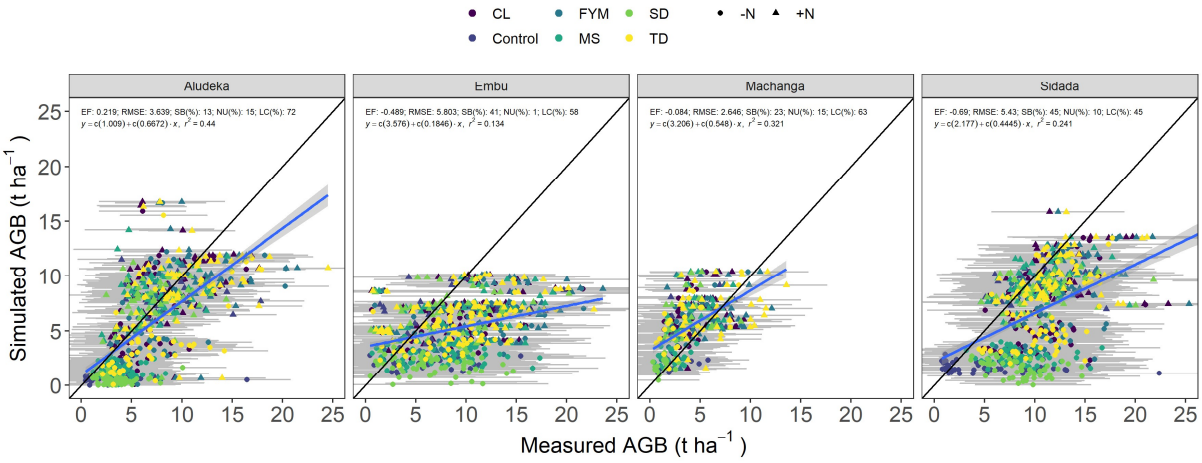
L390-393: Authors can remove this part since calibration is widely recognized to improve model performances.

We think it is important to keep it because the model performance improvement is from leave-one-site-out cross-validation. Hence the performance at each site was improved despite the fact that the calibration was done with only the other three sites. This is notable and indicates that the improvement of DayCent parameters suited the tropical conditions, and was not an overfitting to each site. We however see that the way we did not state clearly enough that most of the results are from the leave-one-site-out cross-validation (e.g., all Figures 4 to 7 are all from this leave-one-site-out cross-validation, despite combining all sites in one graph). We will make this clearer in the next version of the article.

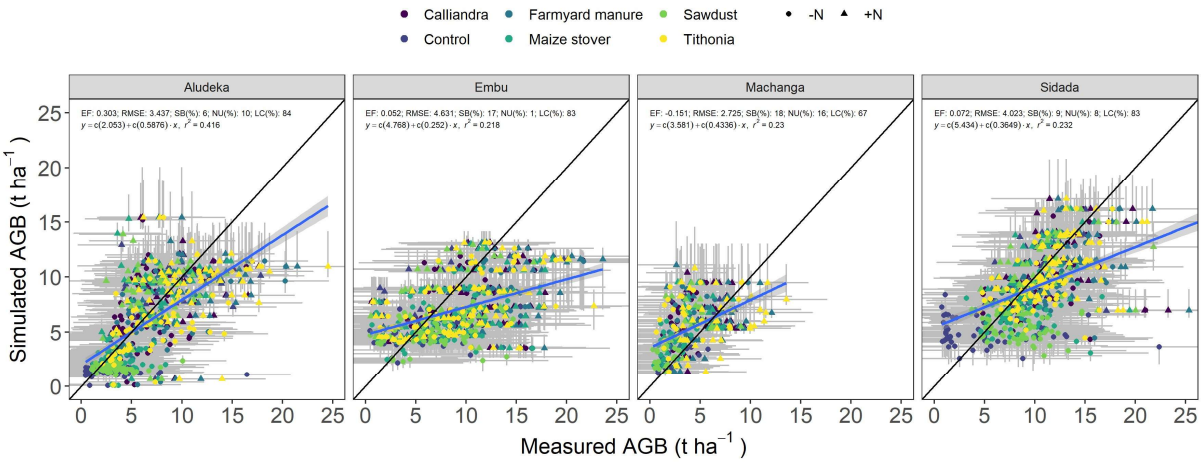
L394: ….and for aboveground biomass for all sites except Machanga. You mean Aludeka?

No, this was 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




Calibrated




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….

Part of this is due to the automatic placement of figures in Latex, and it makes most sense to correct this in the final article. We will put special attention on the chronological order in the overhaul during review.

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 will report the results per site in the main manuscript (replacing Figures 3 and 5). We will also improve Fig. A5 with this in mind (adding evaluation statistics). Further, we will overhaul the Discussion Section after conducting additional simulations, to discuss how well yield and AGB for each ISFM management practice can be simulated at each site.

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.

The reason why we stated that they were simulated reasonably was in consideration of the large uncertainty of measurements of cumulative $N_2O$ (confidence intervals of measurements overlap the 1:1 line) and that the modelling efficiencies were positive. We would thus think that the data we have cannot give a definite answer whether DayCent performs well or not. However, we agree that the simulation is not very good as stated above. Thus, we will remove $N_2O$ from the GHG balance if a model recalibration will not improve the simulations. For now, we 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, that there was no systematic under- or over-prediction of cumulative $N_2O$ emissions, and that simulated $N_2O$ emissions were in the uncertainty range of measured $N_2O$ emissions, does not provide evidence that this current DayCent calibration is not suitable to represent $N_2O$ emissions. This is important for the predictions of the GWP. Because the simulation of SOC change showed low bias, we can conclude that this part of the GWP is well represented. The contributions to GWP between 80% (Aludeka) and 20% of the GWP (other sites; Fig. 8), are less certain. 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 DayCent's simulation of $N_2O$ emissions is superior to using emission factors (dos Reis Martins et al.,