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 aim to refine the discussion part on objective iii, after a model recalibration, based on the comments of reviewer 1 and 2.

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. We will add a map with the locations to the supplement.

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:

to mid-May in Sidada). They were conducted with the static chamber method (Hutchinson and Mosier, 1981). Two measuring frames were permanently installed in the plots for a whole rainy season (one within, one between maize rows). The chambers (0.27 × 0.375 × 0.11 m) were made of polyvinylchloride and equipped with a vent tube and a fan to homogenize gas inside them before gas sampling. Gas samples were extracted through a septum sealed sampling port using a 60mL Polypropylene syringe. Four gas samples were collected within the closure period at interval of 0, 15, 30 and 45min. The pooling method, as described in (Arias-Navarro et al., 2017) was used, where gas samples from the two subplot chambers (inside and between maize rows) were combined per time point and thus mixed in the same syringe. All the samples collected were measured using a SRI 8610C gas chromatography (456-GC, Scion Instruments, Livingston, United Kingdom) equipped with an electron capture detector for $N_2O$ analysis. The fluxes per surface area of the chamber were determined by using the linear slope the gas concentration over time (Pelster et al., 2017; Barthel et al., 2022). The measured $N_2O$ emissions were evaluated at two

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 will add this to the next 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 nonunity slope (<1; overestimation at low values, underestimation at high values). A bias is considered the 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) that we applied are a better approach to assess whether bias exists.

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

Yes, we reformulated: "The changes ranged from an increase of 0.2 t $CO_2$ equivalent $ha^{-1}$ $yr^{-1}$ to reductions of 1 t $CO_2$ equivalent $ha^{-1}$ $yr^{-1}$."

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 will consider these in the overhaul of the manuscript.

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.

We changed the sentence to: "This is also the case for the present study; the results of the model simulations with the initial parameter sets looks good for the absolute SOC stocks, due to limited change, but not for to the changes in the SOC stocks (Fig. A7 vs 5). "

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

We wanted to express it in terms of that can be related to first order kinetics. After your comment we decided to replace it by 2 and 8 as suggested.

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.

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” to the end of the sentence.

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.

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.

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: “However, the strong differences in the yield-scaled GWP between treatments, such as a 72, 32, 63 and 14 % lower yield-scaled GWP in the FYM 1.2+N treatment compared to the control-N treatment, show that ISFM can still lead to an improvement compared to no- or low-input input treatments and that the yield-scaled GWP is highly relevant in practical terms.” Based on the comments of reviewer, we will recalibrate and reconsider GWP and change it to only $CO_2$ emissions.

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

We agree changed it to “positive” only.

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 agree and removed this sentence after reconsideration.

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

We will consider this, but results may change after the recalibration we will do.

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 depth to the Figure A3: (0-20 cm in this version, will be 0-30 in the next). 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 at each plots 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 spotting this.

| 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 spotting this. 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 A6.** Yield response curve of DayCent to varying levels of mineral N application (control + N treatment, without organic resources) using the calibrated DayCent parameters. 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 applied per season in the simulations was evenly split between the 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 are not entirely sure if we will keep figure A10, after the overhaul of the manuscript. But in case we do, we will increase the font size as suggested.