Dear Moritz Laub,

you revised version of the manuscript: "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" has undergone a second round of reviews (3 reviewers). All reviewers are in agreement that the manuscript has improved considerable, with two suggesting some minor revisions including some clarifying questions and adjustments (revs 2 and 3) whereas the 3rd reviewer remains still critical (see the comments from rev 1 this text). This is particularly the case for the prior and posterior distribution which need to be addressed. This should be achievable relatively straight forward. Similarly the individual minor suggestions for the manuscript provided by all three reviewers can be incorporated right away. Following this, I am accepting the manuscript for publication in BG with subject to minor revisions.

with kind regards

Lutz Merbold

Dear Lutz Merbold,

Thank you for acknowledging the changes we made to improve the manuscript from the previous version. We have done our best to address the remaining concerns of the reviewers. We put specific focus on addressing the concerns that reviewer 2 had regarding the new method of initialization with measured SOC pools, the derivation of the coefficients of variation from the prior and the posterior, where reviewer 2 rightly pointed out an oversight from our side. Based on the feedback from reviewer 2 and reviewer 3, we also did calibration once more with wider priors (×1.5), which improved the results even further and led to the results showing a clear distinction between the prior and the posterior. We think that we have addressed all the important concerns of both reviewers with these changes. We hope that with these changes implemented, you will consider the manuscript to be acceptable for publication. Thank you very much for your efforts in handling the manuscript.

Kind regards on behalf of all coauthors,

Moritz Laub

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.

After the revisions made, the paper has strongly improved. All the raised issues were solved, the flow is now clear and figures were made more understundable for readers. Based on all these considerations, the manuscript can be considerd acceptable for publication in its present form.

Thank you for your positive assessment of the revisions that we made.

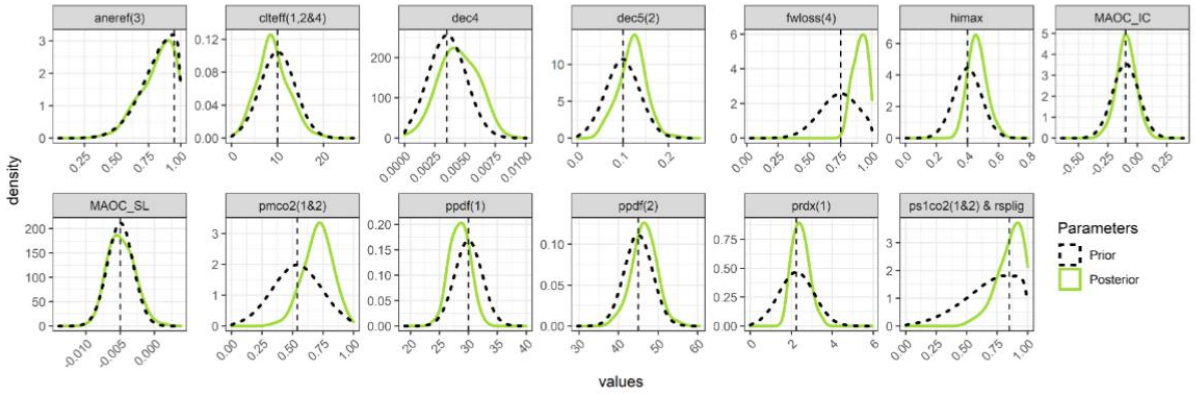
Reviewer 2:

The revision has addressed many of the issues raised in the first review. There has been a significant improvement in the content, flow, and structure, which has increased the readability of the manuscript. In addition, authors have incorporated two significant changes in the methodology section: (1) the selection of prior, and (2) the initialization of SOC pools. Reviews of this newly added section are provided in the subsequent paragraphs. Another area of concern is in the newly updated result section 3.2, which reported the posterior estimates from the inverse modeling using SIR algorithms. Here, the marginal distribution for the posterior is similar to the marginal distribution from the prior, indicating little or no influence of the dataset on the posterior suggesting nothing is learned from the data assimilation exercise. This is contrary to the title's claim of "robust DayCent model calibration …". For these reasons, I recommend a major revision before recommendation for publication. More details are provided below:

Thank you for your feedback. We agree that the original title was not fitting anymore. We were actually using Bayesian calibration to derive model parameters which we can consider to be robust and it is not the calibration that is robust. Therefore, we changed the title to "Modelling integrated soil fertility management for maize production in Kenya using a Bayesian calibration of the DayCent model.

~~A robust DayCent model calibration to assess the potential impact of~~ Modelling integrated soil fertility management ~~on~~ for maize ~~yields, soil carbon stocks and greenhouse gas emissions~~ production in Kenya using a Bayesian calibration of the DayCent model.

Apart from that, we have increased the range of the prior, based on your comment. This improved the simulation outcomes further and the change between prior and posterior is now obvious, making it clear that the data informed the model parameters.



SOC pool initialization: One of the major changes involved replacing the long historical simulation with measured SOC, which was adjusted backward in time and initialized at the start of the Experiment. With the update, the posterior parameter can be defined as p(θ|D,M,MAOM) and now conditions on data (D), model (M) and measured mineral associated organic matter (MAOM). Therefore, any future simulation leveraging this study and aiming to understanding the regional or national GHG balance within SSA, as represented by the four-experiment station, now also requires measured or estimated value of MAOM. Therefore, the simulation approach, which requires MAOM measurement, may limit broader use of the model in SSA region. Furthermore, the initialization of

SOC pools in the process-based ecosystem model may introduce significant bias in the model's estimates of SOC stock changes (Fallon and Smith, 2000, Zhou et. al., 2023). Both methods: (1) long historical simulations to equilibrium and (2) initializing model pools with measurement—have been extensively studied and used in literature. In my personal viewpoint, both methods are equally valid given adequate testing and reasoning. Both methods have strengths and weaknesses. The strength led to higher accuracy, while the weaknesses can introduce significant bias and contributing toward higher uncertainty.

We agree that both approaches have their pros and cons. If there is good knowledge available on the land-use history, a model spin-up may be the better choice. While the model now relies on measured or estimated MAOC, which may introduce its own level of uncertainty, global predictions of MAOC, including for soils in Africa, are available (see Georgiou et al., 2022; specifically Supplementary Figure 16 of their article) and the ranges presented on their maps for the SSA region (i.e., that about 60-90% of SOC is MOAC) agree with our measurements. These maps are produced based on observed MOAC values, extrapolated based on measured gradients of precipitation, temperature, and vegetation. In theory, one could argue that these factors include similar information as a model spin-up, which is based on historical data about land use and includes local weather. In the context of SSA, and in Kenya specifically, we consider that data on land-use history is subject to more uncertainty than the currently available maps of MOAC (e.g, the soilgrids have more than 2000 profiles in Kenya, while a documentation on historical land use is not readily available). This implies that several assumptions have to be made about land-use history, for example by using expert opinion combined with a rules based approach (see e.g., Kamoni et al., 2007). With maps, the only choice to make is which map to use, making this approach more reproducible and favorable in our case. Nonetheless, we agree that it comes with its own uncertainties, which we have addressed in the discussion. Based on the literature you suggested, and our own observations at scale, we recommend to reduce this uncertainty by working with a baseline scenario and an improved scenario instead.

610 ~~In that sense, the~~ Nevertheless, soil property maps are also subject to uncertainty. For example, differences between different SOC maps used in model initialization propagate into differences in the changes in SOC stocks (Zhou et al., 2023). It was shown that uncertainty of the simulated effect of a soil management practice on the difference of SOC stocks compared to a counterfactual is lower than the uncertainty of the simulated temporal development of SOC stocks (Zhou et al., 2023). Therefore, it may be best practice to work with a baseline and an improved scenario. Both spinup and SOC map initialization
615 have their shortcomings and in the end the model user must make an informed decision on which initialization method they consider subject to less uncertainty, based on which data is locally available.

Prior distribution: Another significant change in the updated manuscript is the introducing of a new prior distribution during the SIR step. As a result, the revised manuscript uses two sets of prior distribution: a uniform prior for the global sensitivity analysis, and a Gaussian prior for the SIR step. This is uncommon and generally not accepted in Bayesian inferences, as the prior is considered our initial beliefs about an uncertain parameter before observing any data. Therefore, introducing two beliefs for the same set of parameters in the model is unusual.

We would not call the distribution of the parameter values used in the global sensitivity analysis (GSA) a "prior" in the Bayesian sense and acknowledge that this has been poorly formulated in the previous version of the manuscript. For example, GSA is only constrained by a maximum and a minimum value – and thus any other distribution than uniform is not possible (a poor use of prior knowledge). For us, GSA was simply a preselection procedure to make the Bayesian calibration computationally feasible. Many studies commonly use background knowledge to select which parameters to calibrate and then use Gaussian priors in the calibration process (e.g., Menichetti et al., 2020; Ťupek et al., 2019). Therefore, if we consider the GSA solely as a preselection step, there

appears to be no impediment to subsequently conducting a Bayesian calibration using a Gaussian prior, if knowledge about the model parameters exists. The text of the methods section has been changed, accordingly.

~~distributions were~~ The ranges used for the global sensitivity analysis ~~, with the upper and lower parameter boundaries~~ were centered around the initial parameter value obtained as described above (section 2.3.2). ~~We based the global sensitivity analysis parameter ranges~~ The upper and lower parameter boundaries were based on previous sensitivity analyses (e.g. Necpálová et al., 2015; Gurung et al., 2020), plausible ranges reported in the DayCent manual and variations observed in different maize

Further, our GSA did not make use of the data. In contrast to Gurung et al. (2020), the sensitivity in our study was calculated based on the mean yields and AGB, and end-of-simulation SOC stocks across sites, and not on the mismatch between the simulated and observed values. Hence, we do not see that performing a Bayesian calibration afterward violates the assumption that that Gaussian prior is formulated before observing the simulations compared to the data. Our prior was mainly informed by previous studies (see comments to your next point, below).

The authors selected coefficients of variations ranging between 5% and 30% for DayCent parameters based on the level of range provided in its manual. However, they did not provide details on how these selection for coefficient of variation satisfy one of the requirements for SIR or similar method, which is that the prior range should cover the entire range of the posterior (Galman, 2014). Also, more detail should be provided on the choice for the coefficient of variation to convincing demonstrate that the empirical data suggest that the prior range is adequate enough for the theoretical understanding of these parameters.

Thanks for bringing this point to our attention. We agree that we had insufficiently described this in the methods. In selecting the coefficients of variation, we considered prior knowledge from pervious Bayesian calibration exercises performed on DayCent. Our aim was to choose a coefficient of variation per range level in a way that our prior covered the range that previous Bayesian calibrations on DayCent had as the posterior. Due to your comment, we have now increased the range of the prior to account for the uncertainty of applying DayCent in tropical conditions by increasing the coefficient of variation by a factor of 1.5. We now added a detailed explanation of all this to the text.

To ensure computational efficiency, we used informed Gaussian priors that were centered around the standard parameter values of DayCent, with different coefficients of variation ~~of 0.05, 0.1,~~ based on different observed ranges in previous studies. To make optimal use of existing knowledge about the parameters, the selected coefficients of variation per range were initially based on previous studies that had performed Bayesian calibration of the DayCent model. The coefficients of variation were chosen in a way that the prior from our study covered the whole range of the posterior from previous studies and then multiplied by a factor of 1.5 to account for the additional uncertainty that arose from applying DayCent at tropical sites. The studies of Gurung et al. (2020) and Mathers et al. (2023) were the basis to derive the coefficient of variation for the parameters dec4, dec5(2), clteff(1,2,&4), ps1co2(1&2) & rsplig, and pmco2(1&2). The study of Yang et al. (2021) was the basis for the parameters ppdf(1) and ppdf(2), and the study of (Necpálová et al., 2015, though not being Bayesian) was the basis for the parameters aneref(3) and fwloss(4). For himax and prdx(1), we looked into the default parameters of annual crops in DayCent, to assure that the whole range of values (0.30-0.55, and 1.1-3.5; 1.7-2.5, respectively) was covered by the prior. The final coefficients of variation were 0.08, 0.15, ~~0.25 and 0.3~~ 0.23, 0.38 and 0.45 for parameters with very small, small, moderate, large and very large ranges (Table 1). For the newly introduced parameters, we used ~~larger~~ large coefficients of variation, namely 0.38 for SL_t and 1 for IC_{MAOC}~~and 0.35 for SL~~, the reason for the latter being an initial test, in which $IC_{t\text{-}MAOC}$ was set to -0.3 instead of -0.1, which proved to be too low, but the uncertainty range with a standard deviation of 0.1 proved to be reasonable. Additionally, all parameters were constrained to remain within their physically sensible limits (i.e., not <0 for all and not >1 for those representing fractions).

Posterior distribution: The manuscript calibrated 13 model parameters after conducting a parameter screening using the GSA. In section 3.2 of the results, the posterior was presented in Figure 2. Throughout the text, a single parameter estimate was provided, but it was not specified which statistics (e.g., mean, mode, median, etc.) was presented. The posterior should be summarized with sufficient statistics, such as the mean, standard deviation, and 95% credible intervals. If the single parameter estimates the mean, mode, or median, there is a significant disagreement between the text and the figure for parameters clteff(1,2,&4) and pmco2(1&2). The reported posterior estimates of 19.1 and 0.82 fall well outside the curve region with higher density. I believe this could be a miscalculation or misinterpretation, and should be thoroughly investigated.

Thanks for pointing this out. The presented parameter set is neither mean, mode, nor the median, it is the parameter set from the posterior that had the highest likelihood, based on data from all four sites combined (i.e., not leaving any site out – the final step after cross-validation). While we had stated this in table 2, the description was very brief and did not appear in the main text. We see how this information could be easily overlooked by the reader. Therefore, we now added a better explanation to table 1. Further, we added the requested statistics to table 1:

Table 1. DayCent model parameters and the coefficient of variation used in the calibration. Displayed are parameters considered for calibration due to total sensitivity index > 2.5% (top) and with a total sensitivity index > 1% (bottom). The remainder of parameters (<1%) are not included in this table and can be found in the supplementary (Table A3). The presented calibrated parameter values correspond to the single parameter set with the highest likelihood, which was derived by using the data from all four sites combined. The posterior was also derived by using the data from all four sites combined. Abbreviations: CV, coefficient of variation; SD, standard deviation, 95% CI, 95% credibility interval; Ly, langley.

		Possible ranges		Initial		Calibrated	Posterior		
Parameter	Description	of values	Units	value	CV	value	mean	SD	95% CI
Included in calibration (total sensitivity >2.5%)									
himax	Maximum harvest index for maize	moderate	g g^{-1} (C)	0.40	0.23	0.43	0.46	0.06	0.36-0.59
ppdf(1)	Optimum temperature for growth of maize	very small	°C	30.00	0.08	28.63	28.5	1.67	25.44-31.57
ppdf(2)	Maximum temperature for growth of maize	very small	°C	45.00	0.08	47.1	46.48	3.03	40.01-52.19
prdx(1)	Potential aboveground production of maize	large	g C m^{-2} Ly^{-1}	2.25	0.38	2.62	2.45	0.39	1.86-3.37
clteff(1,2&4)	Tillage multiplier for SOM turnover	large	unitless	10.00	0.38	4.93	9.02	3.35	2.91-15.78
aneref(3)	Min. impact of soil anaerobiosis on SOM turnover	large	unitless	0.95	0.38	0.79	0.82	0.13	0.55-0.99
dec4	Max. turnover rate of passive SOM pool	very large	g g^{-1} yr^{-1}	0.0035	0.45	0.0060	0.004	0.001	0.001-0.007
dec5(2)	Max. turnover rate of slow SOM pool	large	g g^{-1} yr^{-1}	0.10	0.38	0.13	0.12	0.03	0.06-0.17
fwloss(4)	Scaling factor potential evapotranspiration	moderate	unitless	0.75	0.23	0.94	0.9	0.05	0.81-0.99
pmco2(1&2)	C lost as CO_2 with metabolic litter turnover*	large	g g^{-1} (C)	0.54	0.38	0.91	0.71	0.11	0.48-0.91
ps1co2(1&2)&rsplig	C lost as CO_2 with structural litter turnover*	large	g g^{-1} (C)	0.85	0.38	0.77	0.86	0.11	0.61-0.99
IC_{MAOC}	Intercept for fraction of MAOC in slow pool	very large	g g^{-1} (C)	-0.1	1	-0.02	-0.1	0.08	-0.25-0.06
SL_t	Slope for time difference of MAOC measurement	large	g g^{-1} yr^{-1} (C)	-0.005	0.38	-0.006	-0.005	0.002	-(0.001-0.008)
Not included in calibration (total sensitivity <2.5% & > 1%)									
frtc(1)	C allocated to roots at planting, without stress	small	fraction of NPP	0.50	0.15	-	-	-	-
frtc(3)	Time after planting at which maturity is reached	small	number of days	90.00	0.15	-	-	-	-
pramn(1,2)	Min. aboveground C/N ratio at maturity	small	C/N ratio	62.50	0.15	-	-	-	-
hiwsf	Max. harvest index reduction with water stress	moderate	g g^{-1} (C)	0.60	0.23	-	-	-	-
teff(1)	Temperature inflection point (SOM turnover)	moderate	unitless	17.05	0.23	-	-	-	-
varat21&22(2,1)	Min. C/N ratio for material entering slow SOM pool	small	C/N	12.00	0.15	-	-	-	-
basef	Soil water of bottom layer lost via base flow	moderate	fraction H_2O	0.30	0.23	-	-	-	-
N2Oadjust_max	Proportion of nitrified N that is lost as N_2O	large	g g^{-1} (N)	0.015	0.38	-	-	-	-
MaxNitAmt	Maximum daily nitrification amount	large	g N m^{-2}	0.40	0.38	-	-	-	-

*(1 - microbial carbon use efficiency)

And we added text on how the calibrated parameter set was derived into the results section 3.2:

Following the global sensitivity analysis, 13 selected model parameters were calibrated using Gaussian priors which were centered around the initial parameter value, with standard deviations according to the uncertainty ranges (Table 1). ~~The ranges of the prior- and the posterior distributions, using~~ It should be noted that the presented calibrated parameter values in Table 1 correspond to the single best parameter set for all four sites combined (i.e., the parameter set that had the highest likelihood in the case of no cross-validation).

Additionally, we thank the reviewer for bringing to our attention the mismatch between the calibrated values that we reported in Table 2 (& results) and the distributions in the figures. The calibrated parameter set reported in the previous version of the manuscript was in fact a mistake from our side. Due to an oversight, we had taken a wrong parameter set from a pre-test of the new calibration, rather than the one of the final calibration. Hence, the reported values were incorrect. This has now been corrected. We updated this information in the table and the main text, using the values from the newest calibration (see above).

Following the global sensitivity analysis, 13 selected model parameters were calibrated using Gaussian priors which were centered around the initial parameter value, with standard deviations according to the uncertainty ranges (Table 1). ~~The ranges of the prior- and the posterior distributions, using~~ It should be noted that the presented calibrated parameter values in Table 1 correspond to the single best parameter set for all four sites combined (i.e., the parameter set that had the highest likelihood in the case of no cross-validation).

Compared to the range of the prior parameter sets, the ranges of the posterior parameter sets calibrated with data from all four sites changed significantly for the parameters fwloss(4) and pmco2(1&2), had a similar mean value but a more narrow distribution for the parameters IC_{MAOC}, prdx(1), and ps1co2(1&2)&rsplig, and changed slightly for the parameters dec4, ~~were similar . Also the four different posterior distributions from~~ dec5(2), ppdf(1), ppdf(2), and himax (Fig. 2). The posterior parameter sets of the leave-one-site-out cross-validations were largely ~~similar to each other (~~in agreement with each other and with the posterior parameter sets calibrated with data from all four sites. The exception was the parameter pmco2(1&2), which was centered around 0.55 for the case that the Aludeka site was left out and around 0.70 for all other cases (Fig. 2). ~~However, several parameters slightly shifted from their initial values to the best parameter values across~~

The parameter that changed most strongly in the parameter sets calibrated with data from all four sites ~~. The strongest differences between the initial and calibrated values existed for the potential maximum maize productivity per radiation (prdx(1~~was the scaling factor for potential evapotranspiration (fwloss(4); from ~~2.25 to 1.85 g C m^{-2} $langley^{-1}$) , the parameter representing the increase of SOM turnover after tillage (clteff(~~0.75 to 0.94) thereby not including the initial value in the 95% posterior credibility interval (0.81 to 0.99; Table 1). Also the CUE of metabolic litter was reduced (by an increase of pmco2(1~~,~~&2) from 0.54 to 0.91 g g^{-1}) but the initial value was still within the 95% posterior credibility interval (0.48 to 0.91 g g^{-1}). The turnover rates increased for both the slow SOM pool (dec5(2~~,&4) ; from 10 to 19.1) . An increase of the turnover rate of the~~); from 0.10 to 0.13 g g^{-1} yr^{-1}) and the passive SOM pool (dec4; from 0.0035 to ~~0.0056~~ 0.0060 g g^{-1} yr^{-1})~~was partly~~ , which was however counterbalanced by a ~~decrease in the turnover rate of the slow SOM pool (dec5(~~reduction of the effect of tillage on decomposition (clteff(1,2,&4); from ~~0.10 to 0.06~~ 10 to 5) and all three of these parameters contained their initial values in the 95% posterior credibility intervals. The maximum harvest index slightly increased (himax; from 0.40 to 0.43 g g^{-1}~~yr~~) and so did the potential production of maize per unit of light interception (prdx(1); from 2.25 to 2.62 g C m^{-2} $langley^{-1}$). ~~Furthermore the loss of carbon from the metabolic litter pool upon decomposition was significantly increased (pmco2(~~Finally, the optimum temperature for maize growth decreased (ppdf(1~~&~~); from 30 to 28.6 °C), while the maximum temperature for maize growth increased (ppdf(2); from ~~0.54 to 0.82 g g^{-1}). The~~ 45 to 47.1 °C). Of the two parameters that translated measured MAOC into SOC in the passive SOM pool~~were altered in opposite directions (IC_{MAOC} ~~, only IC_{MAOC} was altered (from -0.1 to ~~-0.21~~ -0.02 g g^{-1}~~; and SL_t, from -0.005 to -0.0024~~) but the initial value was still in the 95% posterior credibility intervals(-0.25 to 0.06 g g^{-1}~~yr^{-1}~~). Overall, the parameter correlations in the posterior parameter set across the four sites were ~~minimal, and in no case stronger than 0.2 (~~low for soil carbon related parameters (around 0.4 at maximum), but stronger correlations existed between plant productivity-related parameters (e.g., -0.7 between himax and prdx(1) and 0.58 between ppdf(1) and ppdf(2); Fig. A3).

Some minor comments and corrections:

Line 20-23: The author claim that: “The model performance and the match between the cross-evaluation posterior credibility intervals for different sites indicated the robustness of the model

parameterization and the reliability of the DayCent model for spatial upscaling of simulation." However, the manuscript did not perform a large-scale simulation, and the claim for "spatial upscaling" should be removed or justified.

We agree that this was misleading, and thus refined the sentence to "for the conditions in Kenya".

one parameter. Together with the model performance for the different sites in cross-validation, this indicated the robustness of the DayCent model parameterization and ~~the reliability of the Daycent model for spatial upscaling of simulations. While~~

---

20 its reliability for the conditions in Kenya. While DayCent poorly reproduced daily N_2O emissions ~~were poorly reproduced by~~

Line 23: provide quantitative values (i.e., EF for daily N2O) instead of just mentioning negative value.

Thanks, we did so.

20 its reliability for the conditions in Kenya. While DayCent poorly reproduced daily N_2O emissions ~~were poorly reproduced by DayCent (all EF values were negative~~(with EF ranging between -0.44 and -0.03 by site), cumulative seasonal N_2O emissions were simulated more accurately (EF ranging between ~~0.03 and 0.62~~ 0.06 and 0.69 by site). The simulated yield-scaled GHG

Line 70: The terms "validated" and "evaluated" were used interchangeably throughout the manuscript. For instance, in line 9, "cross-evaluation" is used but in line 164, "cross-validation" is used.

Thanks for pointing this out. We changed this to "evaluation of the model" throughout the text, with the method of evaluation being named "cross-validation". Any "cross-evaluation" was removed.

term "C sequestration" instead of "mineralization of SOC"

The ambiguity of this statement was also pointed out by other reviewers and we changed it accordingly:

soil and optimizing crop yield (that is, sustainable intensification). ISFM can be a source of N_2O to the atmosphere (Leitner
80 et al., 2020) but ~~at the same time mitigate CO~~compared to standard practices, it reduces SOC losses or even increases SOC (Laub et al., 2023a), thereby mitigating CO_2 ~~emissions due to the mineralization of SOC (Laub et al., 2023a)~~emissions.

Line 134: Should this

"Tithonia diversifolia (TD) green manure and Calliandra calothyrsus (CC) prunings, low quality stover of Zea mays (MS) and sawdust from Grevillea robusta trees (SD), locally available farmyard manure (FYM) and a control treatment"

be written as following.

“Tithonia diversifolia (TD) green manure, Calliandra calothyrsus (CC) prunings, low quality stover of Zea mays (MS), sawdust from Grevillea robusta trees (SD), locally available farmyard manure (FYM) and a control treatment”

Thanks! This was in fact not very clear and has been revised:

105 ments, with two crops per year, one in the long rainy season and one in the short rainy season. The experimental design was identical at all four sites and has been described in detail in earlier publications (Chivenge et al., 2009; Gentile et al., 2011; Laub et al., 2023a, b). Organic resource treatments consisted of high quality *Tithonia diversifolia* (TD) green manure~~and~~, high quality *Calliandra calothyrsus* (CC) prunings, low quality stover of *Zea mays* (MS)~~and~~, low quality sawdust from *Grevillea robusta* trees (SD), locally available farmyard manure (FYM), and a control treatment (CT) without organic resource additions.

In Table 1. values for model parameters and coefficient of variation seems truncated given the table description (i.e., parameter values and coefficient of variations were missing)

The table description has been updated to specify that not all parameters considered in the GSA are shown in Table 1.

Table 1. DayCent model parameters and the coefficient of variation used in the calibration. Displayed are parameters considered for calibration due to total sensitivity index > 2.5% (top) and with a total sensitivity index > 1% (bottom). The remainder of parameters (<1%) are not included in this table and can be found in the supplementary (Table A3). The presented calibrated parameter values correspond to the single parameter set with the highest likelihood, which was derived by using the data from all four sites combined. The posterior was also derived by using the data from all four sites combined. Abbreviations: CV, coefficient of variation; SD, standard deviation, 95% CI, 95% credibility interval; Ly, langley.

Equation before line 185, if SOC stock estimates are for 0-30 cm as IPCC-recommended, it should be:

〖SOC〗_30 (kg 〖ha〗^(-1))= (1- β^30)/(1-β^15)* 〖SOC〗_15

Provide the value used for beta^15 and beta^30 used in the equation. The equation number is also missing.

The comment likely refers to the track change version of the article, in which the equation was still visible (in red, indicating that it had been removed in this version).

Because we have removed the equation and went for a different approach to calculate SOC stocks in the first 30 cm of soil, it is not longer necessary to provide beta. This new approach is described in lines 146ff (track-change version) or 139ff (clean version).

Line-261, it is a little confusing and not clear what the author wants to convey. Specifically, data availability and which model parameters and value used for initialization.

The sentences were overhauled:

~~To parameterize the organic inputs, the mean lignin content~~ It was assumed that the organic resource inputs had the same properties across all sites (i.e., mean values of lignin contents and C/N ~~ratio of the different organic materials across sites (~~ratios per organic resource were assumed; Table A2)~~were used~~. This approach was used because measurements were not available for all sites and years, and was justified as an analysis of variance of data from the years 2002, 2003, 2004, 2005 and

Line-266: It was not clear whether the author's discussion about aboveground biomass (AGB), yield (Y), and harvest index (HI) is based on the measured data or modeled values. In DayCent Y = HI*AGB (for grain crops). However, the parameter HIMAX (maximum harvest index) is adjusted due to stress to (HI <= HIMAX).

This was about the measured data. We added this to the sentence, as follows:

This was the mean value of measured grain C content across sites (standard deviation 1.8%) in the short rainy season 2018 and long rainy season 2019 (data not shown). Given the strong correlation between maize grain yield and aboveground biomass in the measured data (r = 0.87), the aboveground biomass data was transformed to harvest index data for the model calibration

Line-395: Please clarify what multiply/divided by 3 and 10 means. Maybe it is self-explanatory when full view of Table-1 is available.

This sentence was reformulated to remove any ambiguity.

For parameters with large and very large ranges, the upper ~~/lower~~ boundaries were the initial parameter values multiplied ~~/~~by 3 and 10, respectively, the lower boundaries were the initial parameter values divided by 3 and 10, respectively. The parameter

Equaton-3: The Likelihood function provided in Equation-3 is applicable to only one type of measurement, such as Yield or SOC. Please provide details on how multiple likelihoods—for SOC, Yield, and Harvest Index were combined, if at all, for the final Bayesian calibration. If they were not combined, please provide an explanation.

In fact, we did combine all types of measurements in the same likelihood function. This was possible by supplying a weighting factor (the inverse of the standard deviation; SD) to the mixed effects model. The formula we used in R for the loglikelihood was:

logLik(lmer(resid~-1+(1|Site/date),weights = (1/SD),data=EC_HI_SOC))

We have further clarified this in the text.

the inverse of the median standard deviation (of each type of measurement at each site) as weight. By using the inverse of the standard deviation of each type of measurement as weight of the zero-intercept model, it is possible to include different types of measurements into the same likelihood function. This is similar to what is done in weighted analyses commonly performed in meta-analyses (Möhring and Piepho, 2009). 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.

Line 571: The posterior credibility intervals in analog to confidence intervals in frequentist statistics and posterior prediction interval analog to prediction intervals. The coverage probability (i.e., 95% of observed) within the 95% Posterior prediction interval is only valid comparison but not with posterior credibility intervals (note that posterior credibility interval < posterior prediction interval).

We agree that these sections were a bit misleading and have removed these comparisons from the text.

were eliminated ~~in Aludeka and Siadada, reduced in~~ at Sidada, reduced at Embu, but increased ~~in~~ at Machanga.

~~The simulated posterior credibility intervals of simulated yields and aboveground biomass contained 50% and 51% of observed data, respectively, showing a that it could not capture the full uncertainty of measurements.~~ While DayCent could

~~Despite the reduction in model performance, the Bayesian calibration effectively captured the uncertainty in SOC stock changes in Aludeka, Embu and Sidada. Overall, 84% of measurements fell within the posterior credibility intervals, though the evaluation was done with~~ -1.9 compared to -4.8 before calibration). DayCent performed well in simulating the variability

Prove all the missing equations used in the analysis, (one such example is the equation for aggregated model output for the GSA) in the supplementary section.

Thanks for this suggestion. We went through the whole manuscript with a focus on this issue and added several equations to the supplementary section.

A1 Pedotranfer functions to derive the hydraulic parameters

The equations used to calculate the soil hydraulic properties were based on the pedotransfer functions of Hodnett and Tomasella (2

745 :

$$\theta_r = 0.22733 - 0.00164 \times Sa + 0.00235 \times CEC - 0.00831 \times pH + 1.8 \times 10^{-5} \times Cl^2 + 2.6 \times 10^{-5} \times Sa \times Cl \quad \text{(A1)}$$

$$\theta_s = 0.81799 + 9.9 \times 10^{-4} \times Cl - 0.3142 \times BD + 1.8 \times 10^{-4} \times CEC + 0.00451 \times pH - 5 \times 10^{-6} \times Sa \times Cl \quad \text{(A2)}$$

$$ln(\alpha) = -0.02294 - 0.03526 \times Si + 0.024 \times SOC - 7.6 \times 10^{-3} \times CEC - 0.11331 \times pH \quad \text{(A3)}$$

$$ln(n) = 0.62986 - 0.00833 \times Cl - 0.00529 \times SOC + 0.00593 \times pH + 7 \times 10^{-5} \times Cl^2 - 1.4 \times 10^{-4} \times Sa \times Si \quad \text{(A4)}$$

750 Here, θ_r, θ_s, α, and n are the soil water retention parameters of van Genuchten (1982), *Sa*, *Si* and *Cl* are Sand, Silt, and Clay content (in %), *BD* is the bulk density (t m^{-3}) CEC is the cation exchange capacity (cmol kg^{-1}), *pH* is the soil pH measured in H_2O, and *SOC* is the SOC content (g kg^{-1}).

The wilting point (WP) and field capacity (FC) values were then calculated as

$$WP = \theta_r + \frac{(\theta_s - \theta_r)}{(1 + (\alpha \times |-15000|)^n)^{1-\frac{1}{n}}} \quad \text{(A5)}$$

755 $$FC = \theta_r + \frac{(\theta_s - \theta_r)}{(1 + (\alpha \times |-330|)^n)^{1-\frac{1}{n}}} \quad \text{(A6)}$$

K_S was calculated using the Saxton and Rawls (2006) equation, with values of the water retention curve, α and n (van Genuchten, 1982) , calculated with the equation from Hodnett and Tomasella (2002):

$$\lambda = \frac{\ln(FC) - \ln(WP)}{\ln(1500) - \ln(33)} \quad \text{(A7)}$$

$$K_S = \frac{1930 \times (\theta_s - FC)^{(3-\lambda)}}{10 \times 60 \times 60} \quad \text{(A8)}$$

760 Here, λ is the slope of logarithmic tension-moisture curve and K_S is the saturated water conductivity (cm s^{-1}).

A2 Equations for the global sensitivity analysis

The means across all sites, which were used in the GSA were calculated as follows:

$$Mean = \frac{1}{n} \sum_{j=1}^{n} \frac{\sum_{i=1}^{N} Mod_{ij}}{N} \quad \text{(A9)}$$

Here n is the number of sites (4), N is the number of modelled values per site, and Mod_{ij} are the individually modelled

765 values. For aboveground biomass and grain yield, N corresponded to the total number of modelled yields and biomass at all treatments and seasons. For SOC and soil N stock N corresponded to the total number of treatments per site. The reason is that because changes in SOC and soil N stocks are expected to be stronger the longer a simulation lasts, only the stocks from the end of the simulation were used.

Reference:

Falloon, P. D., & Smith, P. (2000). Modelling refractory soil organic matter. Biology and Fertility of Soils, 30(5–6), 388–398. https://doi.org/10.1007/s003740050019

Zhou, W., Guan, K., Peng, B., Margenot, A., Lee, D., Tang, J., Jin, Z., Grant, R., DeLucia, E., Qin, Z., Wander, M. M., & Wang, S. (2023). How does uncertainty of soil organic carbon stock affect the calculation of carbon budgets and soil carbon credits for croplands in the U.S. Midwest? Geoderma, 429, 116254. https://doi.org/10.1016/j.geoderma.2022.116254

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014.Bayesiandata analysis, Third edit. ed, BDA3. Boca Raton : CRC Press, Boca Raton.

Reviewer 3:

The study uses a Bayesian calibration approach (sampling importance resampling) with leave -one-site-out cross-validation to calibrate the biogeochemical model Daycent to yields, biomass and SOC at four sites in Kenya. The authors addressed adequately the suggestions of previous reviewers and the community comment and improved the quality of the manuscript. Overall, the manuscript is well-written, methods are sound and described sufficiently, Results are well described and followed by a sensible Discussion. I have some suggestions and comments (see below), and suggest to publish the manuscript after these minor revisions.

Thank you for your overall positive assessment of our manuscript and for the constructive feedback that you provided. We have incorporated the necessary changes, based on your feedback. See the details below.

I refer to the track changes version with my line numbers.

Abstract L33: Daycent is well-suited to estimate the impact of ISFM The impact of ISFM on what? -> Please add

Thanks for spotting this unclear formulation. We added "on maize yields and SOC changes" to the sentence.

25 application of mineral N and of manure at a ~~moderate~~ rate of 1.2t C ha^{-1} yr^{-1}. In conclusion, our results indicate that DayCent is well-suited ~~to estimate~~ for estimating the impact of ISFM ~~,~~ on maize yield and SOC changes. They also indicate that the

Introduction

L82: so a propagation of errors is possible in upscaling exercises

We can be sure the errors propagate in upscaling exercises even if you don't track them, you probably mean: So an estimation of uncertainties is possible in upscaling exercises

You are right, this was not formulated well. Your suggestion was incorporated.

65 long-term experiments. Ideally, this calibration would include the uncertainty in the model parameters and model outputs (Clifford et al., 2014), so ~~a propagation of errors~~ an estimation of uncertainties is possible in upscaling exercises (Stella et al., 2019). This is especially relevant given a recent study showing considerable uncertainty in DayCent's SOM turnover rates,

L103: ISFM can…. but at the same time mitigate CO2 emissions due to the mineralization of SOC

That's an ambiguous formulation, please rephrase to an unmistakable sentence.

Thanks. We rephrased as follows:

soil and optimizing crop yield (that is, sustainable intensification). ISFM can be a source of N_2O to the atmosphere (Leitner
80 et al., 2020) but ~~at the same time mitigate CO~~compared to standard practices, it reduces SOC losses or even increases SOC (Laub et al., 2023a), thereby mitigating CO_2 ~~emissions due to the mineralization of SOC (Laub et al., 2023a)~~emissions.

L105: displaying the confidence in model parameters by Bayesian calibration

Not clear what you mean by that

We reformulated this as follows:

Kenyan conditions using experimental data from four long-term experiments, displaying the ~~confidence in~~ uncertainty of model parameters by Bayesian calibration, and (iii) to use the calibrated model to gain understanding of the GHG balance of the different ISFM treatments.

Methods

L253: ,. taken calculated with the equation

- Typo, remove 'taken'

Removed, thanks for spotting this.

method to estimate K_{*S}, K_{*S} was calculated using the Saxton and Rawls (2006) equation, with values of the water retention curve, α and n (van Genuchten, 1982), ~~taken~~ calculated with the equation from Hodnett and Tomasella (2002). The equations can be found in the supplementary material (A1).

L495: in CO2 eq kg-1 maize grain yield

- in kg CO2 eq kg-1 maize grain yield

We interpreted this comment as a hint to missing units and added also the unit of the annual GHG balance.

Here, ΔSOC is the change in SOC content (kg C ha^{-1} yr^{-1}), N_2O the cumulative N_2O flux (kg N_2O ha^{-1} yr^{-1}). The CH_4 oxidation capacity was not considered, because it usually makes a very limited contribution to GHG balance in rainfed cropping systems (Lee et al., 2020) and we did not have data to evaluate the reliability of this simulated flux. In addition to the net annual GHG balance (in t CO_2eq ha^{-1} yr^{-1}), we calculated the yield-scaled GHG balance (in CO_2eq kg^{-1} maize grain yield) by dividing the cumulative GHG balance over the entire simulation period by cumulative simulated yields (dry matter base).

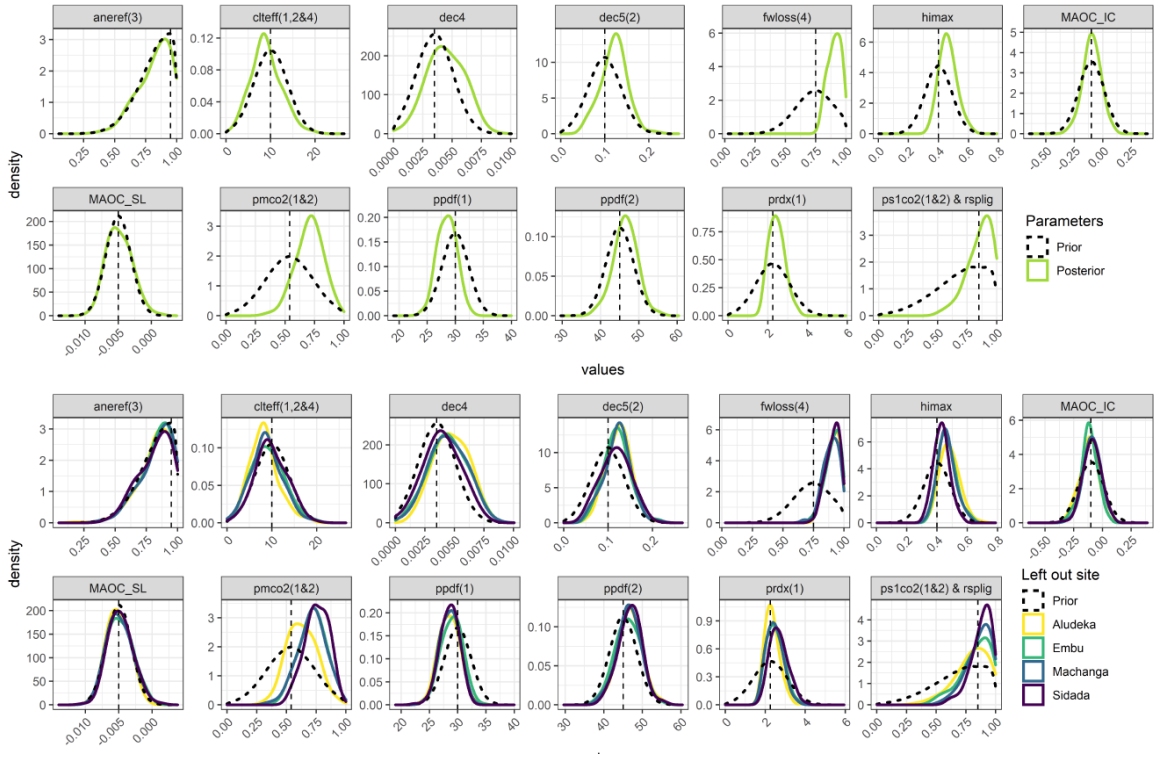
Results

Figure 2

My visual impression is that prior and posterior distributions are quite similar.

Why is the posterior less narrow in Figure 2 compared to the prior? Wouldn't one expect the calibration to constrain the parameters and give a narrower posterior compared to the prior?

We agree that they were very similar. Based on your comment and a comment from reviewer 2, we therefore increased the range of the prior to by increasing the coefficient of variation by a factor of 1.5. This led to the data clearly constraining the posterior. We have updated the results accordingly.


aneref(3)
clteff(1,2&4)
dec4
dec5(2)
fwloss(4)
himax
MAOC_IC
MAOC_SL
pmco2(1&2)
ppdf(1)
ppdf(2)
prdx(1)
ps1co2(1&2) & rsplig
density
values
Parameters
Prior
Posterior

aneref(3)
clteff(1,2&4)
dec4
dec5(2)
fwloss(4)
himax
MAOC_IC
MAOC_SL
pmco2(1&2)
ppdf(1)
ppdf(2)
prdx(1)
ps1co2(1&2) & rsplig
density
Left out site
Prior
Aludeka
Embu
Machanga
Sidada

Following the global sensitivity analysis, 13 selected model parameters were calibrated using Gaussian priors which were centered around the initial parameter value, with standard deviations according to the uncertainty ranges (Table 1). ~~The ranges of the prior- and the posterior distributions, using~~ It should be noted that the presented calibrated parameter values in Table 1 correspond to the single best parameter set for all four sites combined (i.e., the parameter set that had the highest likelihood in the case of no cross-validation).

Compared to the range of the prior parameter sets, the ranges of the posterior parameter sets calibrated with data from all four sites changed significantly for the parameters fwloss(4) and pmco2(1&2), had a similar mean value but a more narrow distribution for the parameters IC_{MAOC}, prdx(1), and ps1co2(1&2)&rsplig, and changed slightly for the parameters dec4, ~~were similar . Also the four different posterior distributions from~~ dec5(2), ppdf(1), ppdf(2), and himax (Fig. 2). The posterior parameter sets of the leave-one-site-out cross-validations were largely ~~similar to each other (~~in agreement with each other and with the posterior parameter sets calibrated with data from all four sites. The exception was the parameter pmco2(1&2), which was centered around 0.55 for the case that the Aludeka site was left out and around 0.70 for all other cases (Fig. 2). ~~However, several parameters slightly shifted from their initial values to the best parameter values across~~

The parameter that changed most strongly in the parameter sets calibrated with data from all four sites ~~. The strongest differences between the initial and calibrated values existed for the potential maximum maize productivity per radiation (prdx(1~~was the scaling factor for potential evapotranspiration (fwloss(4); from ~~2.25 to 1.85 g C m^{-2} langley^{-1}) , the parameter representing the increase of SOM turnover after tillage (clteff(~~0.75 to 0.94) thereby not including the initial value in the 95% posterior credibility interval (0.81 to 0.99; Table 1). Also the CUE of metabolic litter was reduced (by an increase of pmco2(1~~,~~&2) from 0.54 to 0.91 g g^{-1}) but the initial value was still within the 95% posterior credibility interval (0.48 to 0.91 g g^{-1}). The turnover rates increased for both the slow SOM pool (dec5(2~~,&4) ; from 10 to 19.1) . An increase of the turnover rate of the~~); from 0.10 to 0.13 g g^{-1} yr^{-1}) and the passive SOM pool (dec4; from 0.0035 to ~~0.0056~~ 0.0060 g g^{-1} yr^{-1})~~was partly~~ , which was however counterbalanced by a ~~decrease in the turnover rate of the slow SOM pool (dec5(~~reduction of the effect of tillage on decomposition (clteff(1,2,&4); from ~~0.10 to 0.06~~ 10 to 5) and all three of these parameters contained their initial values in the 95% posterior credibility intervals. The maximum harvest index slightly increased (himax; from 0.40 to 0.43 g g^{-1}~~yr~~) and so did the potential production of maize per unit of light interception (prdx(1); from 2.25 to 2.62 g C m^{-2} langley^{-1}). ~~Furthermore the loss of carbon from the metabolic litter pool upon decomposition was significantly increased (pmco2(~~Finally, the optimum temperature for maize growth decreased (ppdf(1~~&~~); from 30 to 28.6 °C), while the maximum temperature for maize growth increased (ppdf(2); from ~~0.54 to 0.82 g g^{-1}). The~~ 45 to 47.1 °C). Of the two parameters that translated measured MAOC into SOC in the passive SOM pool~~were altered in opposite directions (IC_{MAOC} ~~, only IC_{MAOC} was altered (from -0.1 to ~~-0.21~~ -0.02 g g^{-1}~~; and SL_t, from -0.005 to -0.0024)~~ but the initial value was still in the 95% posterior credibility intervals(-0.25 to 0.06 g g^{-1}~~yr^{-1}~~). Overall, the parameter correlations in the posterior parameter set across the four sites were ~~minimal, and in no case stronger than 0.2 (~~low for soil carbon related parameters (around 0.4 at maximum), but stronger correlations existed between plant productivity-related parameters (e.g., -0.7 between himax and prdx(1) and 0.58 between ppdf(1) and ppdf(2); Fig. A3).

Figure 2 caption: Not clear what you want to say by 'uncertainty-based Bayesian model calibration', but since this is not a term generally used or a method description, I would leave out the term 'uncertainty-based'.

We changed this formulation as follows:

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). The uncertainty ranges of the priors were based on the range of parameter values found in the literature and increased by a factor of 1.5, because DayCent was applied to tropical site, while historically, it was mostly calibrated based on temperate sites. Dashed vertical lines represent the values of the initially selected parameter set. The posterior distributions are based on all four study sites combined. For the description of the parameters see Table 1.

Figure 7: ‘the black solid line the simulation by the best parameter set for each site’ You did not calibrate by site, but the caption can be understood as if you did. Since the panels are per site anyway, I would recommend to omit ‘for each site’ here in the caption.

Thanks for spotting this ambiguity. We omitted “for each site” as suggested:

Figure 7. Measured (dots) versus simulated SOC stocks over time at the four study sites for the different organic resource and chemical nitrogen fertilizer treatments. Error bars represent 95% confidence intervals for measured data, the black solid line the simulation by the best parameter set~~for each site~~. Grey bands represent the 95% credibility intervals of the model posterior simulations, calibrated by leave-one-site-out cross-validation. Note that due to intense soil erosion, data from Machanga was not used in the calibration process.

Figure 8: Credibility intervals for cumulative fluxes are quite narrow, and do not cover the 1:1 line. Are these really credibility intervals? Unlike the other figures, N2O was not calibrated. I think they are quite misleading here, since N2O was not included in the calibration so of course they remain narrow if you put narrow posterior distributions. Or is it variance that is displayed? Please add explanation in the caption.

You are right and we added this fact to the explanation:

Figure 8. Simulated compared to measured N_2O emissions at the four study sites for the different organic resource and chemical nitrogen fertilizer treatments, based on the calibrated parameter set using leave-one-site-out cross-validation. Displayed are the measured versus modelled per treatment for the days where measurements were conducted (top) and for the mean of cumulative flux measurements per season using the trapeziod method (bottom). The 808 data points (top) correspond to the daily measurements from the experimental treatments over one to two seasons, depending on the site. Symbols represent the different organic resource and chemical nitrogen fertilizer treatments. Error bars represent 95% confidence intervals (measurements) and credibility intervals (simulations). Note that the credibility intervals are only informed by yield, SOC and harvest index data and therefore do not represent the full uncertainty of N_2O emissions. Abbreviations: EF, Nash-Sutcliffe ~~modeling~~ model efficiency; RMSE, root mean squared error; SB, squared bias; NU, non-unity slope; LC, lack of correlation.

For claiming that the posterior distributions are suitable for upscaling this must also be true for N2O, while my view for N2O a realistic uncertainty estimate is not shown.

This is true, we added a sentence on this in the results and another one in the discussion:

per season, there was a better agreement between the simulated and measured values. All sites, except Machanga, showed positive model efficiencies (highest in Embu, 0.62; lowest in Sidada, 0.03~~;~~), but generally underestimated the uncertainty around cumulative N_2O emissions (Fig. 8). Additionally, the correlation between simulated and measured N_2O emissions was

655 Nonetheless, the fact that the uncertainty around predicted cumulative N_2O emissions was lower than the uncertainties of the measurements indicates that the posterior, which was only calibrated with yield, SOC, and harvest index data, underestimates the uncertainty around N_2O emission predictions.Thus, although DayCent’s simulations of N_2O emissions are superior to using emission factor approaches (dos Reis Martins et al., 2022), simulating N_2O emissions remains challenging and highly uncertain due to the complexity of the processes involved and their high temporal and spatial variability. Given the limited bias

Which ISFM method is simulated with highest accuracy etc?

If you target a robust fit for upscaling the effect of different ISFM methods, then it might be worth presenting the bias and rmse per treatment across site.

Thanks for this suggestion. We agree and have added this to the Supplementary Section.

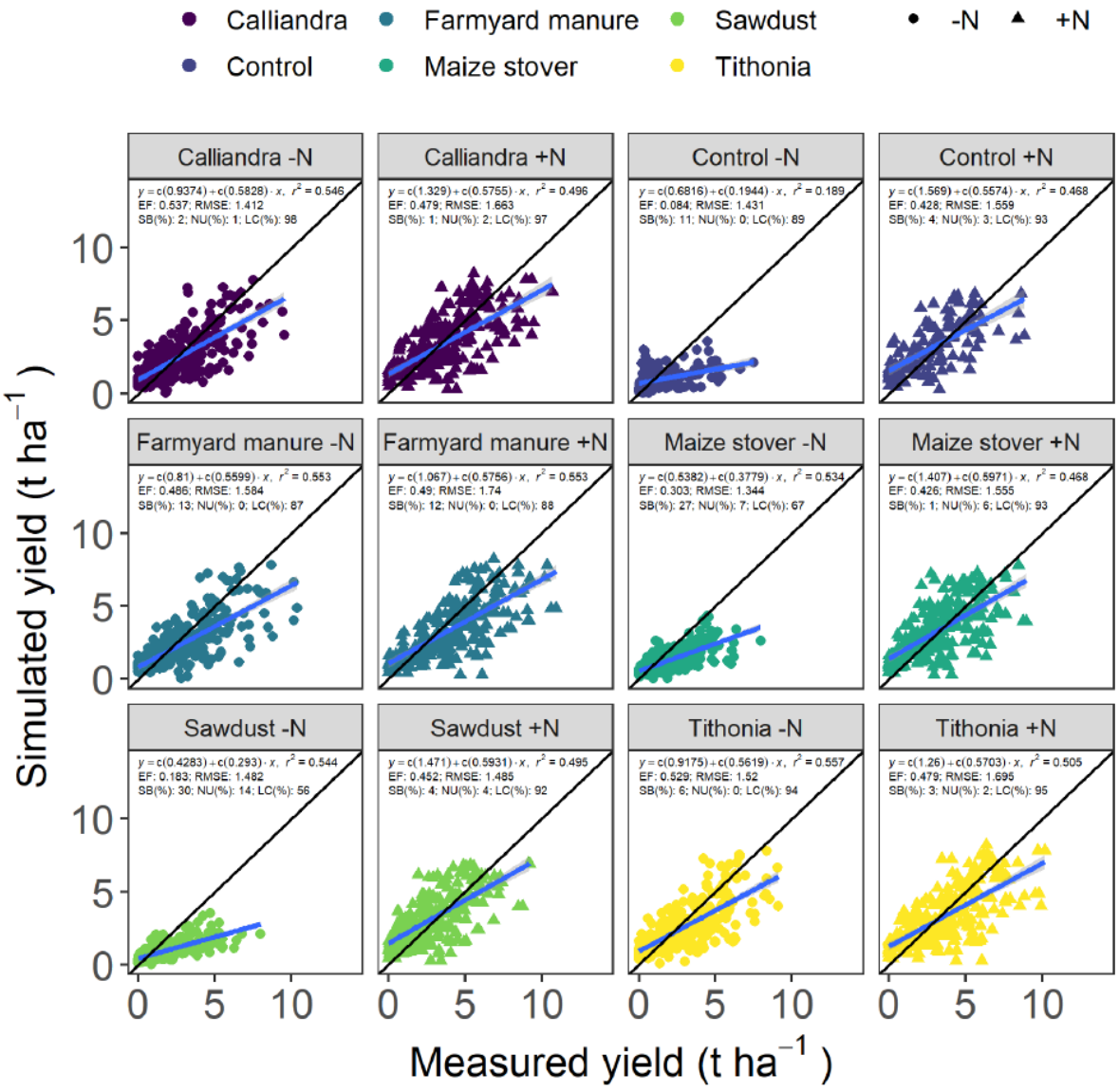


Figure A7. Treatment-specific simulated compared to measured maize grain yields at the four study sites for the calibrated parameter set by leave-one-site-out cross-validation. Abbreviations: EF, Nash-Sutcliffe model efficiency; RMSE, root mean squared error; SB, squared bias; NU, non-unity slope; LC, lack of correlation.

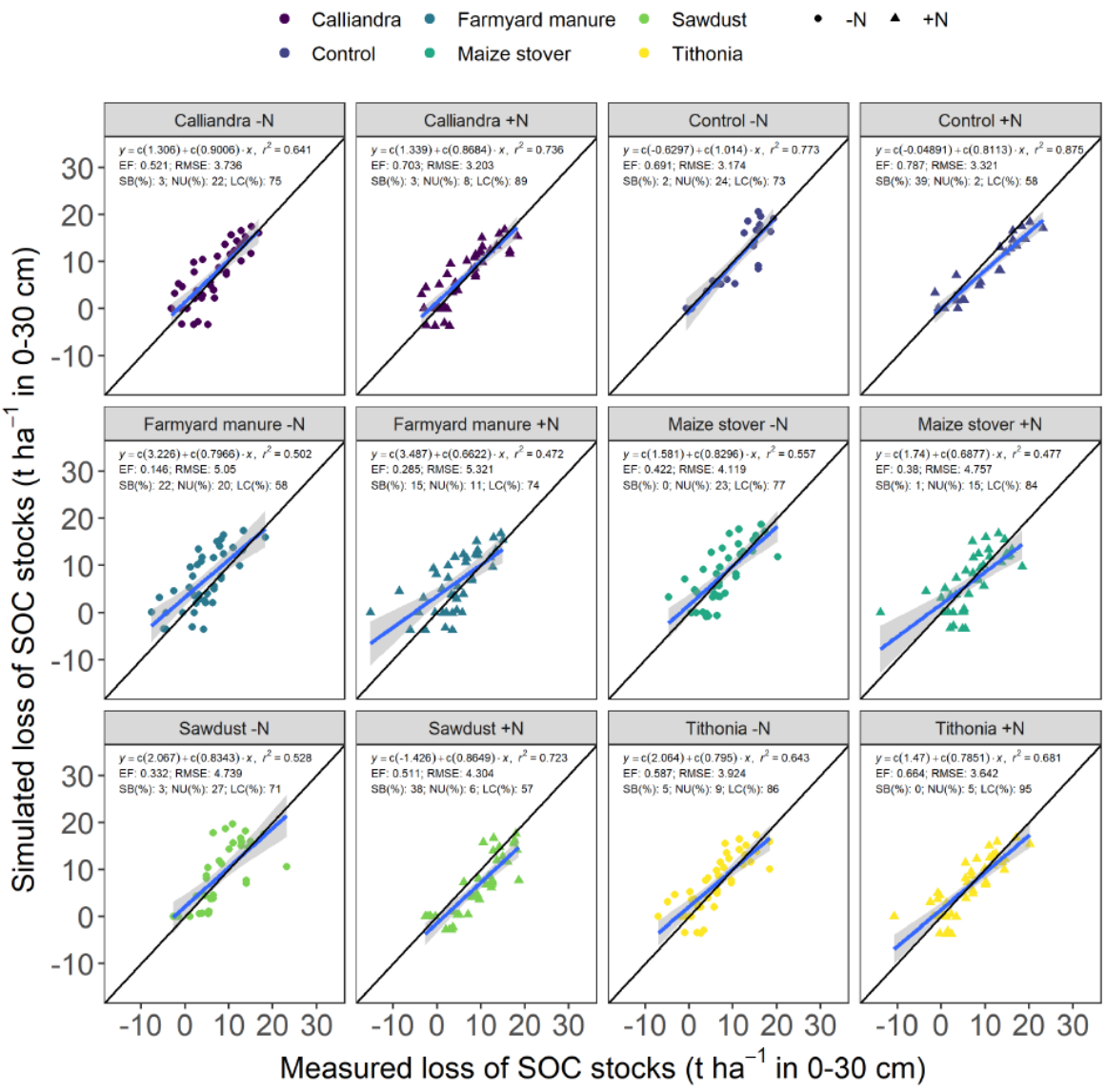


Figure A8. Treatment-specific simulated compared to measured changes in SOC stocks (without the Machanga site) since the start of the experiment at the four study sites for the calibrated parameter set by leave-one-site-out cross-validation. Abbreviations: EF, Nash-Sutcliffe model efficiency; RMSE, root mean squared error; SB, squared bias; NU, non-unity slope; LC, lack of correlation.

We further added a few sentences addressing these new results in the results section:

+N treatments. Nonetheless, DayCent was able to acceptably simulate the variability of grain yields across sites by organic resource and mineral N treatment (model efficiencies between 0.30 and 0.54; with values for control -N (0.08) and sawdust -N (0.18) being the exception; Fig A7). Interestingly, DayCent poorly distinguished the mean yields and aboveground biomass of treatments with high compared to very high rates of N inputs (i.e., the differences between the different organic resources and the control within the +N treatment). An additional test of the model sensitivity of mean yields to different levels of mineral N

~~the evaluation was done with~~ -1.9 compared to -4.8 before calibration). DayCent performed well in simulating the variability of the changes in SOC stocks across sites, evaluated by organic resource and mineral N treatment (also computed without Machanga). With the exception of the ~~site that was not used in calibration.~~ treatments farmyard manure ±N, maize stover ±N, and sawdust -N model efficiencies were between 0.51 and 0.79 (with RSME between 3.2 and 4.3 t ha^{-1}; Fig A8) with the highest performance for the control +N (0.79) and control -N (0.69) treatments. The other treatments still had positive model efficiencies (0.15 to 0.42), but the SOC losses of the farmyard manure treatments were overestimated (EF of 0.15 for -N, 0.29 for +N, RSME of 5.1 and 5.3).

And in the discussion section:

able for upscaling of model simulations. Specifically, the yields of the ISFM treatments applying farmyard manure, *Calliandra*, and *Tithonia* were simulated well, both with and without the addition of mineral N fertilizer (Fig A7). The changes in SOC stocks for the control, *Calliandra*, and *Tithonia* treatments were also simulated well across sites, while DayCent underestimates the SOC buildup from farmyard manure treatments (Fig A8). However, on should keep in mind that the season-to-season yield
570 variability is captured less accurately than the mean yields (lower RMSE) and that changes in SOC are better represented at

Figure 9: Please explain 9b in the caption (Mention 9 a b c in the caption.)

Thank you. We added this to the caption:

Figure 9. Cumulative simulated greenhouse gas (GHG) balance of N_2O emissions and CO_2 emissions due to loss of SOC at the four study sites for different organic resource and chemical nitrogen fertilizer treatments ~~,~~ combined throughout the simulated period (16 years for Aludeka/Sidada; 19 years for Embu/Machanga). Displayed are the GHG balance a) per area of land and year, b) the difference of GHG balance per area of land and year to a no-input treatment, and c) the yield-scaled GHG balance. The GHG balance is expressed in CO_2 equivalent over a 100-year horizon.

In several table & figure captions you explain the lowercase letters:

Same lowercase letters indicate the absence of a significant difference in XYZ …. Easier to read would be a positive formulation: Different lowercase letters indicate a significant difference in XYZ between …

We agree that the positive wording you suggest sounds simpler, but it is ambiguous and strictly speaking not correct (see Piepho, 2018). To make it simpler, we adjusted it to the formulation that Piepho (2018) suggested: “Means not sharing any letter are significantly different”.

Table A2. Mean measured chemical characteristics (and 95% confidence intervals) of organic resources applied at all sites. Measurements were available from Embu and Machanga from 2002 to 2004, all sites from 2005 to 2007 and in 2018. Significant differences in residue properties were found between the different organic resources, but not between sites and years. ~~Same letters within the same~~ Mean values in a row ~~indicate the absence of significant differences for that property~~ not sharing any lowercase letter are significantly different from each other ($p < 0.05$). Abbreviations: n.c. = not classified * according to Palm et al. (2001). The table is adopted from Laub et al. (2023a) under the creative common license 4: http://creativecommons.org/licenses/by/4.0/.

Figure A2. Subsoil SOC stocks for the 2.5-4.7 kt ha^{-1} equivalent soil mass layer, corresponding to an approximate soil depth of 15-30 cm. Displayed are the least square means estimated by the linear mixed model described in (Laub et al., 2023a) for planted plots by treatment (left) and site (right). Error bars display the 95% confidence intervals. ~~Same lowercase letters indicate the absence of a significant difference in SOC stocks between treatments~~ Mean values at ~~the same~~ each site not sharing any lowercase letter are significantly different from each other (left figure)~~or between sites (~~. In the right figure~~;~~ , mean values per site not sharing any lowercase letter are significantly different from each other (all $p < 0.05$). Abbreviations: CC, *Calliandra*; CT, control; FYM, farmyard manure; MS, maize stover; SD, sawdust; TD, *Tithonia Diversifolia*. 0, 1.2 and 4 correspond to C additions of 0, 1.2 and 4 t C ha^{-1} yr^{-1}.

Discussion & Conclusion

These sections make sense to me and I have no further comments.

Thank you for your constructive feedback.

References for the revision:

Georgiou, K., Jackson, R.B., Vindušková, O., Abramoff, R.Z., Ahlström, A., Feng, W., Harden, J.W., Pellegrini, A.F.A., Polley, H.W., Soong, J.L., Riley, W.J., Torn, M.S., 2022. Global stocks and capacity of mineral-associated soil organic carbon. Nat Commun 13, 3797. https://doi.org/10.1038/s41467-022-31540-9

Gurung, R.B., Ogle, S.M., Breidt, F.J., Williams, S.A., Parton, W.J., 2020. Bayesian calibration of the DayCent ecosystem model to simulate soil organic carbon dynamics and reduce model uncertainty. Geoderma 376, 114529. https://doi.org/10.1016/j.geoderma.2020.114529

Kamoni, P.T., Gicheru, P.T., Wokabi, S.M., Easter, M., Milne, E., Coleman, K., Falloon, P., Paustian, K., 2007. Predicted soil organic carbon stocks and changes in Kenya between 1990 and 2030. Agriculture, Ecosystems & Environment, Soil carbon stocks at regional scales 122, 105–113. https://doi.org/10.1016/j.agee.2007.01.024

Menichetti, L., Kätterer, T., Bolinder, M.A., 2020. A Bayesian modeling framework for estimating equilibrium soil organic C sequestration in agroforestry systems. Agriculture, Ecosystems & Environment 303, 107118. https://doi.org/10.1016/j.agee.2020.107118

Piepho, H.-P., 2018. Letters in Mean Comparisons: What They Do and Don't Mean. Agronomy Journal 110, 431–434. https://doi.org/10.2134/agronj2017.10.0580

Ťupek, B., Launiainen, S., Peltoniemi, M., Sievänen, R., Perttunen, J., Kulmala, L., Penttilä, T., Lindroos, A.J., Hashimoto, S., Lehtonen, A., 2019. Evaluating CENTURY and Yasso soil carbon models for CO2 emissions and organic carbon stocks of boreal forest soil with Bayesian multi-model inference. European Journal of Soil Science 70, 847–858. https://doi.org/10.1111/ejss.12805