“In their article, ‘Comparing the MEMS v1 model performance with MCMC and 4DEnVar calibration methods over a continental soil inventory’, Viskari et al. compare two different parameter optimisation methods, Markov Chain Monte Carlo (MCMC) and the 4-Dimensional Ensemble Variational data assimilation method (4DEnVar), to optimise the MEMS v1 model.

Using SOC and carbon fraction data (POM vs MAOM) from the LUCAS 2009 soil inventory, the authors calibrated selected model parameters for 322 soil samples for which the POM and MAOM fractions were known, and analysed how these parameters influence steady-state SOC projections for 17,430 other LUCAS data points. The study includes a twin experiment (to assess if the algorithms were able to find the correct parameter set), two calibration scenarios using different assumptions about the fraction of net primary production entering the soil, and a large-scale validation.

The authors report that both calibration approaches produce similar results despite yielding different parameter sets and a different distribution of simulated SOM between POM and MAOM. They also explore how NPP-related assumptions alter calibration outcomes. Their results highlight the sensitivity of SOC model calibration to litter input assumptions and the implications of parameter differences for projected POM and MAOM distributions across Europe.

Obtaining parameter values through calibration for large amounts of data can be a computationally very costly procedure, as pointed out by the authors. Therefore, the evaluation of different methods to obtain suitable parameter values more efficiently is a valuable effort that can speed up parameterisation in the future. A strong point of the manuscript is that it does not only describe positive results, but also focusses on the pitfalls of model calibration, such as obtaining different parameter values that perform equally well, or a different simulated distribution of SOM among different simulated pools. These aspects of the model calibration process are often ignored in the literature and making modellers aware of these is highly important to advance this field.

The manuscript is well written and the results are clearly presented, although I missed a more quantitative evaluation of the calibration and validation results. I think it is an important contribution to the field of SOC development, which is regularly confronted with limitations when large amounts of data need to be used for model calibration. I hope my feedback can improve the quality of the manuscript, and make some aspects more clear to the readers.

Throughout my feedback, I mention certain published articles. These have been chosen based on their scientific relevance, and I leave it up to the authors whether they want to include these in their manuscript or not.”

Our gratitude for the generally positive view of the content of the manuscript as well as the nuanced and detailed recommendations regarding how to improve. We hope that

we have sufficiently addressed your points and feel that as a whole they have strengthened the manuscript. Additionally, the given references were much appreciated with majority of them now included in the manuscript.

As a general note, the line numbers given in the responses refer to the track-changes version manuscript.

"My main feedback is the following:

- Throughout the manuscript the authors use the wording 'to a meaningful degree', without specifying what this means. This should be done, as it seems this term is used with the same meaning as 'significantly different', but it is not clear which criteria the authors use when applying this term."

As the primary author, my apologies on this one as I did use the term a lot more than was reasonable. It was at times trying to avoid using significantly in a situation where weren't talking about the results of a statistical test, but kind of got out of hand. We have now gone through the manuscript and either removed the term or been clearer with it.

- "One of the main outcomes of the manuscript is that the different calibration methods resulted in similar model outcomes (Fig. 3; although there are some notable differences as shown in Fig. 4) with different parameter values (Fig. 2). In addition, the projections show clear differences in the distribution of SOC between POC and MAOC (Fig. 7). Both are a clear example of equifinality, an important but often overlook concept in SOM modelling, and environmental modelling in general. I encourage the authors to have a look at this concept, and include this in their manuscript as it is highly relevant to interpret their results. The following articles could be used as a guide: Sierra et al. (2015; https://doi.org/10.1016/j.soilbio.2015.08.012), Marschmann et al. (2019; https://doi.org/10.1016/j.envsoft.2019.104518), Beven et al. (2006 ; https://doi.org/10.1016/j.jhydrol.2005.07.007), Van de Broek et al. (2025, https://doi.org/10.5194/bg-22-1427-2025), and Luo et al. (2017 ; https://www.jstor.org/stable/26155933)"

Excellent point. We are aware of equifinality and that is the reason we performed the twin experiment in the first place as it is an important assessor of equifinality. With that written, though, we should have referenced the issue by name as it does provide context to the issues discussed in the manuscript. Additionally, once again our gratitude for the provided references.

We added a small part to the Introduction section where we bring up equifinality and how to even be able to determine if that is an issue requires multiple calibrations. Now, from line 78, it reads:

For example, equifinality is a known issue in ecosystem modelling, where there are multiple parameter sets that produce a similar model output (Sierra et al., 2015; Marschmann et al., 2019). Establishing if this is affecting the model system under study requires repeating the calibration multiple times which is prohibited by too heavy calibration approaches.

Then we expanded the paragraph in the Discussion section mentioned also in the line-by-line comments to highlight how the different parameters produced by the calibration are an example of equifinality. We do, however, also explain that the reason this is surprising to us is how the parameter sets are tied to the calibration method used as, for instance, the MCMC never resolved the calibration in the same part of the parameter space than the 4DEnVar calibration. The paragraph on line 571 now goes:

What is striking, though, is how much the parameter sets produced by the two calibration methods in both litter distribution scenarios differ from each, even with the higher f_{doc}, they perform approximately equally well with regard to the total SOC measurements in the validation dataset. As mentioned in the Introduction, equifinality, a situation where there exists multiple parameter sets that produce similar model outputs, is a known issue in ecosystem modelling and is evidently represented by the results here. The notable element here is that the calibration method itself determines the resulting parameter set as even when repeated, the MCMC calibration approach does not suggest the solution is in the same part of the parameter space as the 4DEnVar results indicate. Generally, twin experiments are efficient first pass to test for equifinality and the challenge can be addressed by reducing the amount of parameters being calibrated, but here there are questions how much those efforts can be relied on in assessing equifinality.

- "Similarly, the parameters selected to be optimized are likely to be 'not-identifiable', meaning that different combinations of these parameters can lead to a similar model output (as observed by the authors). The authors would have been able to draw stronger conclusions about the comparison between the calibration techniques if only 'identifiable parameters', with only one solution, would have been optimised. I encourage the authors to discuss the implications of this, for example using the articles mentioned in the previous point, in addition to Guillaume et al. (2019; https://doi.org/10.1016/j.envsoft.2019.07.007) and Lam et al. (2022; https://doi.org/10.1016/j.matcom.2022.03.020)"

While the 'identifiable parameter' is an important concept in theoretical discussion and modelled systems that can be well-observed, we would argue that with ecosystem models, and especially SOC models, it is very challenging to meet the identifiable parameter threshold. Which is, again in our view, why this test is so rarely applied in the field of ecosystem modelling.

For example, take our current study here. Our measurements are of two distinct types (SOC and MAOM:SOC ratio), are across multiple ecosystems and have considerable

amount of noise. We must make very strong assumptions regarding the driver data and have to assert a steady state situation which is an unfortunate necessity. Furthermore, as already discussed in the manuscript, there is conflict between the two different measurement types.

To be able to meet the identifiable standards in this work we would either have to reduce the model structure or limit the scope of the application. Both are valid arguments, but that is not the focus of the work here.

Since this hits so close to the equifinality topic that we addressed in the previous point, we could not figure out a way to introduce this topic in the manuscript without it coming across as either repetitive or a side path that's not immediately clear in its relevance.

- "The description of the 4DEnVar method is very technical, and difficult to understand for a non-expert. As the difference between this method and MCMC is a core aspect of the manuscript, I would encourage the authors to include a paragraph where the 4DEnVar method is explained in layman terms, with the differences with MCMC being highlighted."

We have added the paragraph requested to the start of the 4DEnVar section. On line 266:

Instead of iteratively exploring the variable space like MCMC does, 4-Dimensional Ensemble Variational data assimilation (4DEnVar) uses an ensemble of model runs with different variable sets and that are independent of each other. The ensemble of model runs is used to approximate information required by other calibration techniques, such as the gradient of the cost function and a mapping from variable space to observation space. Because there is no need for a large amount of model run repetitions such as in MCMC, this method is a computationally much faster. However, this approach is built on certain assumptions – in particular that the observations can be predicted by a linear combination of the different ensemble members - which make it important to test before-hand how well it is able to find the correct values in different systems

- "The discussions and conclusions would benefit from a quantitative description of both calibration and validation results for both methods using multiple error metrics, which is currently lacking. As a result, the reader currently has to rely on only the plots to interpret the results."

We are in complete agreement with this request, and this was honestly something we should have included in the original manuscript.

In the Results section, we added a new table that contains both the RMSE and mean error values as an indicator of bias to line 489:

	f_{doc} 0.15	f_{doc} 0.35
MCMC	42.5 / 27.4	31.3 / 7.4
4DEnVar	29.8 / -1.9	32.0 / 14.2

Table 4: The error statistics for the different parameterizations with regard to the validation dataset. The first value is for the root mean square error (RMSE) and the second for the mean error (ME). The unit for all the values is t C ha-1.

We also rewrote the preceding paragraph to reflect these values and correct a previously mistaken interpretation of the error behaviour. The new paragraph begins on line 478:

To examine the impact of the new parameter sets, Figure 3 presents the differences between the measurements and model projections across all the validation sites, while Table 4 shows both the Root Mean Square Error (RMSE) and mean error (ME) representing bias in regard of the validation dataset for each parameter set. While the 4DEnVar parameter sets produces a somewhat symmetric error distribution around zero in both calibrations, with the higher f_{doc} there is a slight apparent tendency towards positive errors. In contrast, the MCMC error distribution shows a notable lean towards positive errors for the lower f_{doc}, while with the higher f_{doc}, the bias is much reduced. Since the SOC errors here are calculated as the measurement minus the model projection, this means that positive errors reflect the parameter set systematically underestimating the SOC projections. It is notable that with the higher f_{doc}, the RMSE values for the two parameterizations are very closer to each other even with the larger positive bias of the 4DEnVar method.

Line-by-line comments:

"General: one of the main parts of the manuscript is the assessment of how the portion of NPP serving as C inputs affects model parameters and performance, but this is not mentioned in the abstract. I would encourage the authors to do so, so this is clear to the reader from the start."

This was an oversight on our part, thank you for pointing this out. While expanding the abstract to also this part of the manuscript, we also condensed some of the abstract to reduce the amount of the characters to account for the additions.

Now, starting from line 12, the abstract reads as:

Abstract. An abundant amount of different data is required to calibrate soil organic carbon (SOC) models to represent ecosystems at large-scale. However, due to challenges related to model state projections, this calibration becomes very computationally heavy with traditional calibration methods. Here, we test 4-Dimensional Ensemble Variational data assimilation (4DEnVar) method to parameterize the MEMS v1 SOC model using data from the LUCAS network and compare its performance against MCMC calibration. Additionally, we performed an experiment where we adjusted the litter input partition to see if the two calibration methods react differently to the change. The total SOC projections from both parameterizations showed similar improvements though the produced parameter sets differed. A thorough analysis revealed that the detailed SOC states differed from each other, but we also lacked information to determine which parameter set was closer to the truth. Furthermore, changing the litter input partition highlighted how much that assumption affects the calibration results with both methods. Our results here establish 4DEnVar as an applicable calibration method for SOC models but also highlight the need for more nuanced validation methods, as well as careful examination on how different data sets affect the model calibration.

"L 32-33: This understanding has not been 'recently advanced', as SOM fractionation is a practice that has been well-established for over three decades (see, for example, Cambardella et al. (1992; https://doi.org/10.2136/sssaj1992.03615995005600030017x))"

This was a bad phrasing from us as it was meant to imply that SOM fractionation has recently been used more in model development. We have reworded this part as well as added the reference listed here to the following starting from line 39:

To provide more nuanced SOC measurements, separating the bulk soil into SOC fractions (Cambardella and Elliot, 1992; Lavallee et al., 2019; Yu et al., 2022), notably the mineral-associated (MAOM) and the particulate organic matter carbon (POM), has been utilized more in current field campaigns. However, though there are different methods…

"L 40-41: instead of mentioning only two such models, it would be worthwhile to acknowledge that many similar non-linear models exist (see, for example, Chandel et al. (2023; https://doi.org/10.1029/2023JG007436) and Le Noë et al. (2023; https://doi.org/10.1038/s43247-023-00830-5))"

We expanded this part slightly to better indicate how the models mentioned are intended as just examples out of many starting from line 46:

To this purpose, numerous models of varying complexities have been developed (Chandel et al., 2023; Le Noë et al., 2023) with different approaches and focuses. Some are simple first-order dynamic models such as RothC (Coleman and Jenkins, 1996) while others are more complicated non-linear models such as MIMICS (Wieder et al., 2014) and Millennial (Abramoff et al., 2022).

"L 53-55: That is correct, but a solution to this problem is to simulate 14C and evaluate this against measurements of d14C, so that both the stocks and turnover times are simulated correctly."

While 14C is an important tool in evaluating SOC models turnover rates, and we are grateful for being reminded of it here, we did not completely understand this comment. If 14C as a constraint, which is a valuable resource, the calibration would still be affected by the assumptions made in the model structure. To give an example, MEMS has surface decomposition pools while in a model like Millennial, and this is just to name a model, all the NPP are directly inputted into soil pools. Thus, even with the inclusion of the 14C data, the structure still impacts the results. Additionally, there is an argument to be made that the proper use of the 14C data requires a layered SOC model that has its own challenges.

None of this is to dismiss this comment and we have added to the manuscript to address this point, rather to explain why we don't present it as a definite answer to the challenge. With this change, the manuscript now reads from line 63:

While there are valuable additional measurement datasets such as 14C (Brunmayer et al., 2024) that can provide important additional constraints for determining effective litter inputs, even these are still affected by how the NPP input is presented to start with in the model.

"L 78-79: this sentence needs more explanation to be understandable by the reader"

Based on this and feedback from another reviewer, we have changed the paragraph as a whole a lot in order to make the benefits of the 4DEnVar method more apparent. Now it is, starting from line 82:

As a more practical alternative to the costly MCMC approach, four-dimensional ensemble variational data assimilation (4DEnVar; Liu et al., 2008) is a novel data assimilation approach, where a model ensemble generated by varying the parameters/variable states of interest is used to determine the optimal parameter and/or state variables. It has already been used for parameter calibration (Douglas et al., 2025; Pinnington et al. 2020) and is much faster than the traditional MCMC methods. It is based on the Four-dimensional Variational data assimilation (4DVar; Le Dimet and Talagrand, 1986), where a model projection is compared with observations and the new initial state for the next iteration is generated from this information. A key difference between MCMC and 4DVar based methods is that the latter use gradient descent methods to determine the next state instead of randomly sampling. While 4DVar has initially been used more commonly for state data assimilation, for example, in weather forecast (Huang et al., 2009), it has also been successfully applied to calibrate ecosystem models (e.g. Raoult et al., 2016; Peylin et al., 2016; Pinnington et al. 2016). However, to implement 4Dvar with observations from multiple different times, an adjoint version of the model is needed which imposes its own challenges and limitations on the application (Thepaut and Courtier, 1991). The 4DEnVar method, however, uses the ensemble to sidestep this requirement by simultaneously running multiple simulations with different parameter sets instead of an iterative solution. When tested in a synthetic experiments, was more effective in determining the correct parameter than the original 4DVar method (Beylat et al., 2025).

"L 97: it seems the model is applied to 20 cm, as the LUCAS data contains data down to 20 cm. Please explicitly state this in the manuscript."

We added a few sentences to the model description section to address this. From line 168 onwards:

The model dynamics represents the depth of the soil measurements used to calibrate it. As we are using the LUCAS data here which is from the top 20 cm of the soil, the resulting MEMS model will thus simulate the SOC dynamics of top 20 cm layer as well.

"Fig. 1: it would be interesting to see where the 322 data points that were used for calibration were located. Can these be highlighted?"

We have added a panel to Figure 1 that shows the calibration datapoint distribution in the LUCAS dataset. The new figure is below.

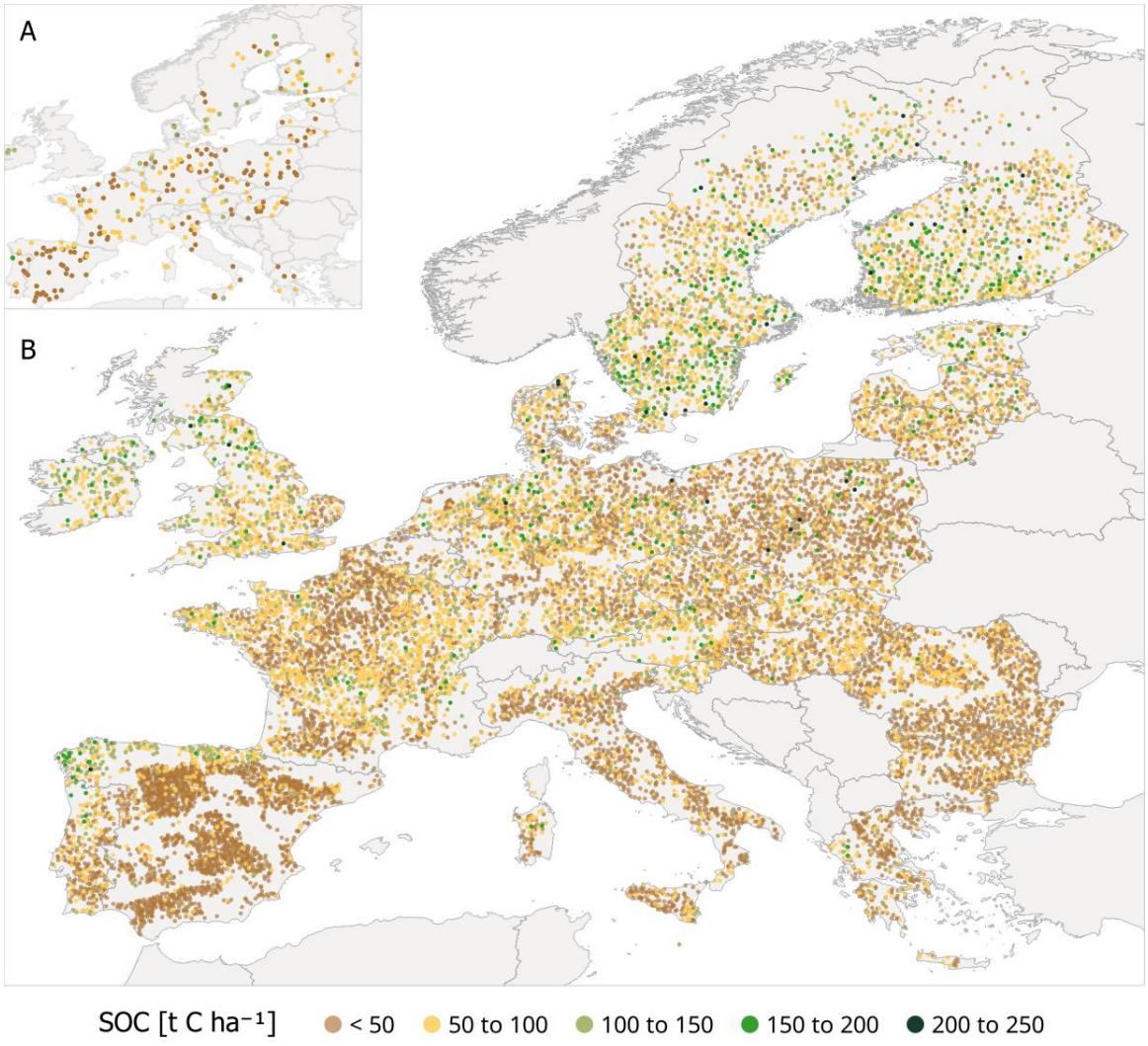
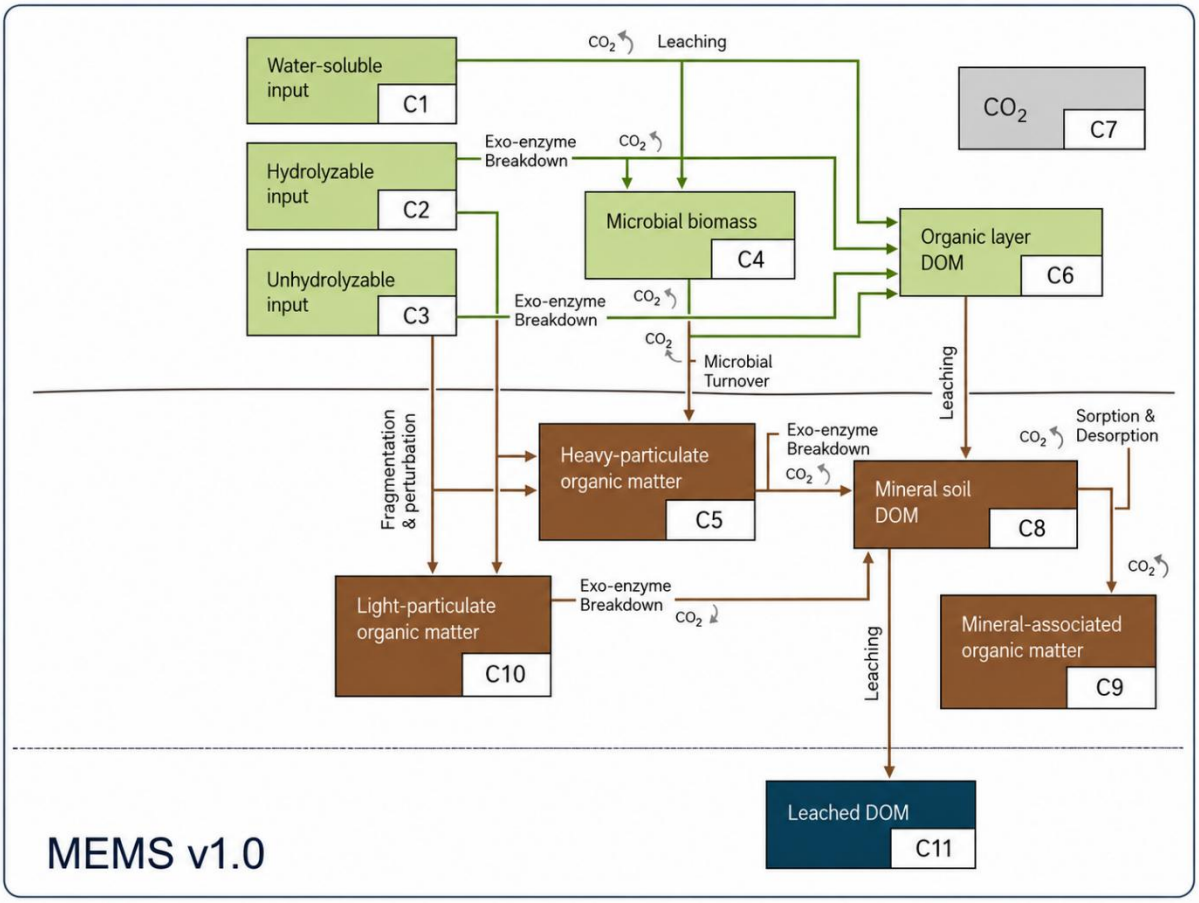


Figure 1: The LUCAS 2009 sampling points across Europe and their SOC stock used for A) calibration and B) validation

"L112: it would be interesting for the reader to see the model structure of MEMS. Perhaps put a graph showing this in the supplement?"

We have added the model structure flowchart as Supplemental Figure 1 and referenced it in the model description. The new figure is below:



Supplemental Figure 1: The MEMS model structure based on the work presented in Robertson et al. (2019).

"L 119: 'were considered in this work': what does that mean? Please clarify."

To improve the clarity, we rewrote the sentence starting from line 172 as

Since the parameterization focuses on the SOC stock, only the model equations affecting MEMS pools C5 (Heavy particulate organic matter), C8 (Dissolved organic matter), C9 (Mineral associated organic matter (MAOM)) and C10 (Light particulate organic matter) were calibrated here.

"L122: please mention for which land use these default parameters were obtained. Are they readily applicable to your simulated forest, grassland and cropland ecosystems?"

In the referenced article the parameters in question had been selected to be representative of the LUCAS network in question. They were not parameterized in that article, but we chose to rely on them as they were a part of the published parameter set for the MEMS model. We have expanded the sentence on line 178 to be clearer about this choice.

Therefore, we used the default parameters values established in Robertson et al. (2019) for the surface processes as they had been chosen in that work as applicable values for the LUCAS network environment.

"L124-127: these equations are very difficult to understand given the generic names of the carbon pools, and the lack of a graph showing the model structure and flows of C between the simulated pools. I suggest to authors to improve this."

As mentioned in the previous response, we have now added the model structure figure in the supplemental material.

"L131-132: a couple words of explanation on the STANDCARB model are needed for readers not familiar with this model to understand."

We have expanded the temperature model reference with a quick description on line 190:

In this work, T_{mod} is the same for all pools and follows the STANDCARB 2.0 model (Harmon et al., 2009) which is an expanded version of the traditional Q10 temperature model where the limiting impact of the high temperatures is accounted for

"L 145: please explain what you mean with 'prior values'. Does this have the same meaning as the prior in a Bayesian calibration?"

Yes, this was meant to refer to the prior establishment for the Bayesian calibration. Upon rereading, we also realized that it was difficult to understand and now have rewritten this sentence to be more explicit from line 205 onwards:

As will explained in Section 2.5, we do need an expected value for these parameters in order to create a prior uncertainty distribution. We chose this value by randomly drawing a parameter value from near the middle of the set of the boundary conditions after testing that the model runs remained stable with these parameter values.

"Table 1: (1) it would be more intuitive for the reader if the pool names (C5, C8, etc.) would be replaced by names of the pools such at POC, DOC, etc. As it is now, this table is difficult to interpret by readers not familiar with the MEMS model. (2) Please clarify what the minimum and maximum values are. (3) Please mention the units of the values. (4) What is meant with the baseline values?"

We have changed the table on line 210 to address all these requests:

Name	Symbol	Expected value	Minimum value	Maximum value
Decomposition rate for heavy particle	k_5	0.0008	0.0001	0.002

organic matter Pool (C5; day^{-1})				
Decomposition rate for dissolved soil organic material pool (C8; day^{-1})	k_8	0.001	0.0001	0.01
Decomposition rate for mineral associated matter pool (C9; day^{-1})	k_9	0.000025	0.00001	0.00004
Decomposition rate for light particle organic matter pool (C10; day^{-1})	k_{10}	0.0005	0.0001	0.0004
Saturation intercept	SC_{Icept}	10.0	5	20
Saturation slope	SC_{Slope}	0.25	0.1	0.4

Table 1: The calibrated parameters chosen for calibration, their assigned expected parameter values as well as boundaries that constrain the lowest and highest values that the parameters are allowed be given during the calibration.

"L151: also here, a graph of the conceptual model of MEMS would help the reader understand how litter inputs are distributed among the model pools."

The conceptual MEMS model chart is now included as Supplemental Figure 1 and referenced at the start of the model description section.

"L 154-157: also here, the equations are not straightforward to interpret because of the use of C1, C2, etc. Better would be to use pool names that are understandable for the reader."

The equations and associated pool names presented are consistent with how they are named in the referenced Robertson et al. (2019) article where the MEMS v1 model was introduced and how those pools are named in the actual code. Thus, we are hesitant to change the naming of the pools here as that would break the shared naming approach across the different sources.

We hope that the addition of the conceptual model figure addresses this concern. Additionally, based on other reviewer feedback we have added a description of the other pools to the paragraph from line 172 and hope that it further aids with better understanding what the different pools represent:

Since the parameterization focuses on the SOC stock, only the model equations affecting MEMS pools C5 (Heavy particulate organic matter), C8 (Dissolved organic matter), C9 (Mineral associated organic matter (MAOM)) and C10 (Light particulate organic matter) were calibrated here. The vegetation decomposition pools C1 (hot-water soluble), C2 (acid soluble) and C3 (acid insoluble) as well as the surface microbial pool (C4) and the dissolved organic matter (C6) do determine the litter input entering to soil C pools, those mechanics were

not included in the calibration as the type of data required to constrain them was not available. Therefore, we used the default parameters values established in Robertson et al. (2019) for the surface processes as they had been chosen in that work as applicable values for the LUCAS network environment. Meanwhile the released CO_2 (C7) and the leached dissolved material to the soil (C11) are cumulative removal pools and do not have any parameters to be calibrated.

"Table 2: it would be good to also explain in the caption what fsol, flig and fdoc are, so the table is understandable by itself"

We added the explanations to the table itself on line 234 as we felt that was the easiest way to represent them

	NPP fraction (r^{eco})	Hot water extricable fraction (f_{sol})	Acid insoluble fraction (f_{lig})	Cold water extricable fraction (f_{doc})
Woody grassland	0.67	0.35	0.15	0.15
Pure grass	0.51	0.35	0.15	0.15
Sporadic grassland	0.59	0.35	0.15	0.15
Cropland	0.43	0.35	0.15	0.15
Mixture	0.77	0.375	0.295	0.15
Broadleaf	0.68	0.4	0.27	0.15
Conifer	0.78	0.35	0.32	0.15

Table 2: The fraction of NPP that is used for litter input and how it is divided into different litter compounds

"L 200: this section is very technical and difficult to understand for a non-expert. I encourage the authors to start this section with a paragraph that explain in simple words how this method works, and how it differs from MCMC. As this is central to your study, it is important that readers can understand how this method works."

We have added the requested paragraph to the start of the section starting from line 267:

Instead of iteratively exploring the variable space like MCMC does, 4-Dimensional Ensemble Variational data assimilation (4DEnVar) uses an ensemble of model runs with different variable sets and that are independent of each other. The ensemble of model runs is used to approximate information required by other calibration techniques, such as the gradient of the cost function and a mapping from variable space to observation space. Because there is no need for a large amount of model run repetitions such as in MCMC, this method is a computationally much faster. However, this approach is built on certain assumptions – in particular that the observations can be predicted by a linear combination of the different ensemble members - which make it important to test before-hand how well it is able to find the correct values in different systems.

"L 277: the approach of performing all optimizations separately for different values of fdoc needs more explanation for the reader to understand why this was necessary"

We have separated this to a new paragraph on Line 371 and added an explanation for the experiment in general. It should be noted that there was no intent to experiment with f_{doc} specifically, rather it was just a test case:

As a part of the testing here, we also wished to experiment how varying assumptions regarding model drivers affected the potential differences between the calibration resultsFor our test case study on the impact of the NPP assumptions on the parameterization, we repeated the calibrations with a small adjustment. We changed the f_{doc} value of grass- and croplands from 0.15 to 0.35. This increases the amount of the litter that is directly deposited to the soi and consequently adsorbed by the mineral matrix instead of being lost during the transition between the surface and soil carbon pools. The logic behind this is that, in our expert opinion, there is a higher proportion of exudates and root litter (i.e. low molecule weight compounds that can directly sorbed by the soil minerals) entering to the topsoil in grasslands and herbaceous crops compared to forests. Thus, this change is suitable for a plausible change to the NPP assumptions and makes an ideal test study to see how it affects the parameterization results and if the system depicted by the parameterizations still remains consistent after the potential change.

"L282-284: something seems to be wrong with this sentence, it is not clear."

On reread, we agree that it was an obscure sentence due to being too long and missing a crucial word. Now on line 376, we have rephrased it to hopefully reflect our intent better:

In our expert opinion, it is likely that there will be a larger proportion of exudates and root litter inputted to the topsoil in grasslands and herbaceous crops to the litter pools compared to forests. Thus, this change is suitable for a plausible minor change to the NPP assumptions and makes an ideal test study to see how it affects the parameterization results.

"L 285-286: please find a better way to mention the initial size of the state variables, perhaps in a table in the supplement."

Moved this to Supplemental Table 1.

"L 340-341: is there an explanation why in the twin experiment, the algorithm found the same parameter values for both optimization methods, while this was not the case when the real data were used?"

This is a very good question. Our theory is that when we produce the synthetic measurements for both the total SOC and the MAOM fraction, those values are internally coherent within the model reality. However, with the measurements, as we make note of in the article, there is a conflict between the measured total SOC and MAOM fractions with the latter being much higher than what suits the model dynamics.

Thus, when calibrating with the real data, the differing results from the two models, again based on our current hypothesis, is due to how they then solve the balance

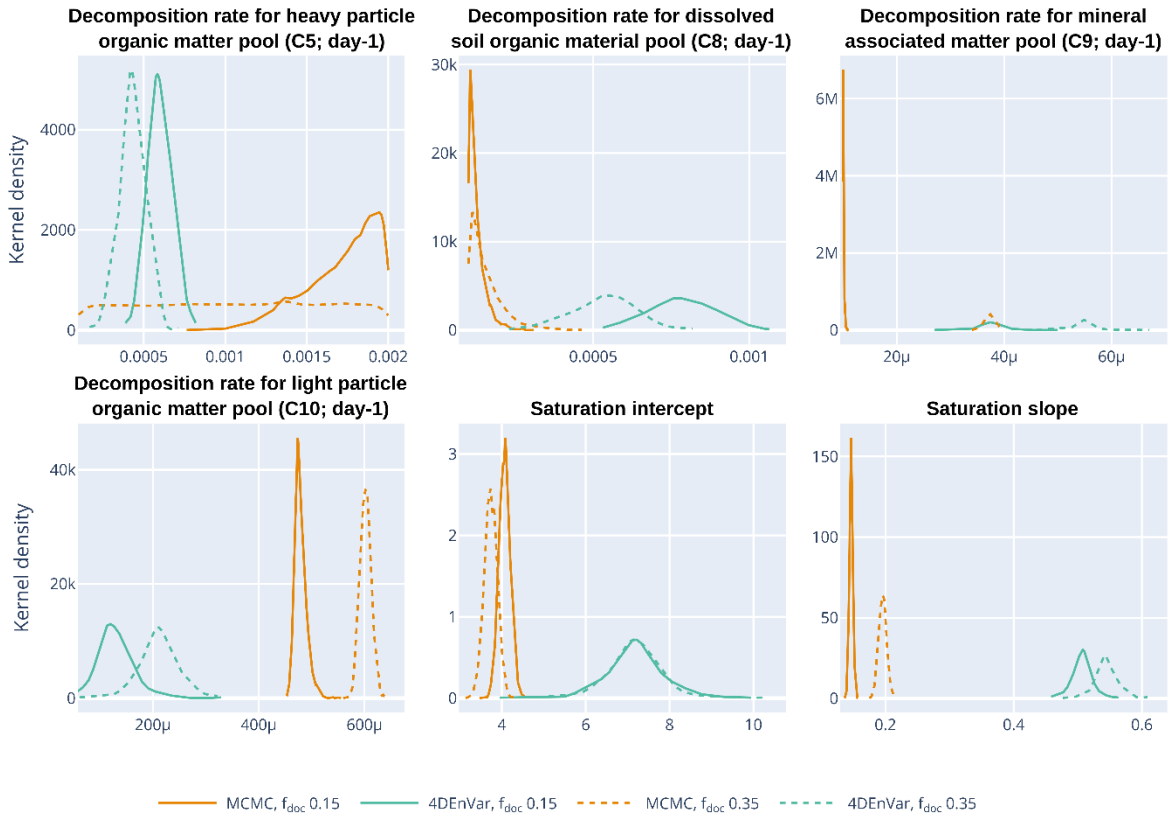
between those two datasets. However, with the synthetic dataset that is not necessary, which allows both methods to just find the same correct parameter set.

We added a sentence on line 589 of the Discussion section to draw attention to this:

While we are not certain of what is driving these systematic differences between calibration sets, we hypothesize that one crucial component is that the total SOC and MAOM fraction measurements appear to incentivize contradicting model behaviours. Our twin experiment results support this theory as, with synthetic datasets, were able to retrieve the same parameter set of both total SOC and MAOM that are internally coherent with the model dynamics.

"Figure 2: Please use more informative names for the parameters. As it is now, names as k5, k8 etc. are not intuitive for the reader and they will not be able to interpret this plot without going back to the methods section."

The figure titles have now been changed as recommended and the new Fig 2 is below:



"L 348: please clarify what you mean by 'expected values'"

We have replaced expected values with statistically likeliest parameter values.

"L 349: please clarify what you mean by 'differ meaningfully'. What criterion do you use for this? Please do so throughout the manuscript where this expression is used."

We changed this to state that they differ with each other more than would be explained by their associated uncertainties. Additionally, we have gone through the manuscript and considered all the parts where the term meaningfully was used.

"L 359-361: this sentence is very difficult to understand, please clarify"

We have attempted to clarify the sentence on line 472:

While there was variance in the produced parameter sets, they overall remained within the uncertainty distribution for any single estimation.

"Table 3: (1) what do you mean by 'expected values'? (2) What are the 'baseline parameters'?"

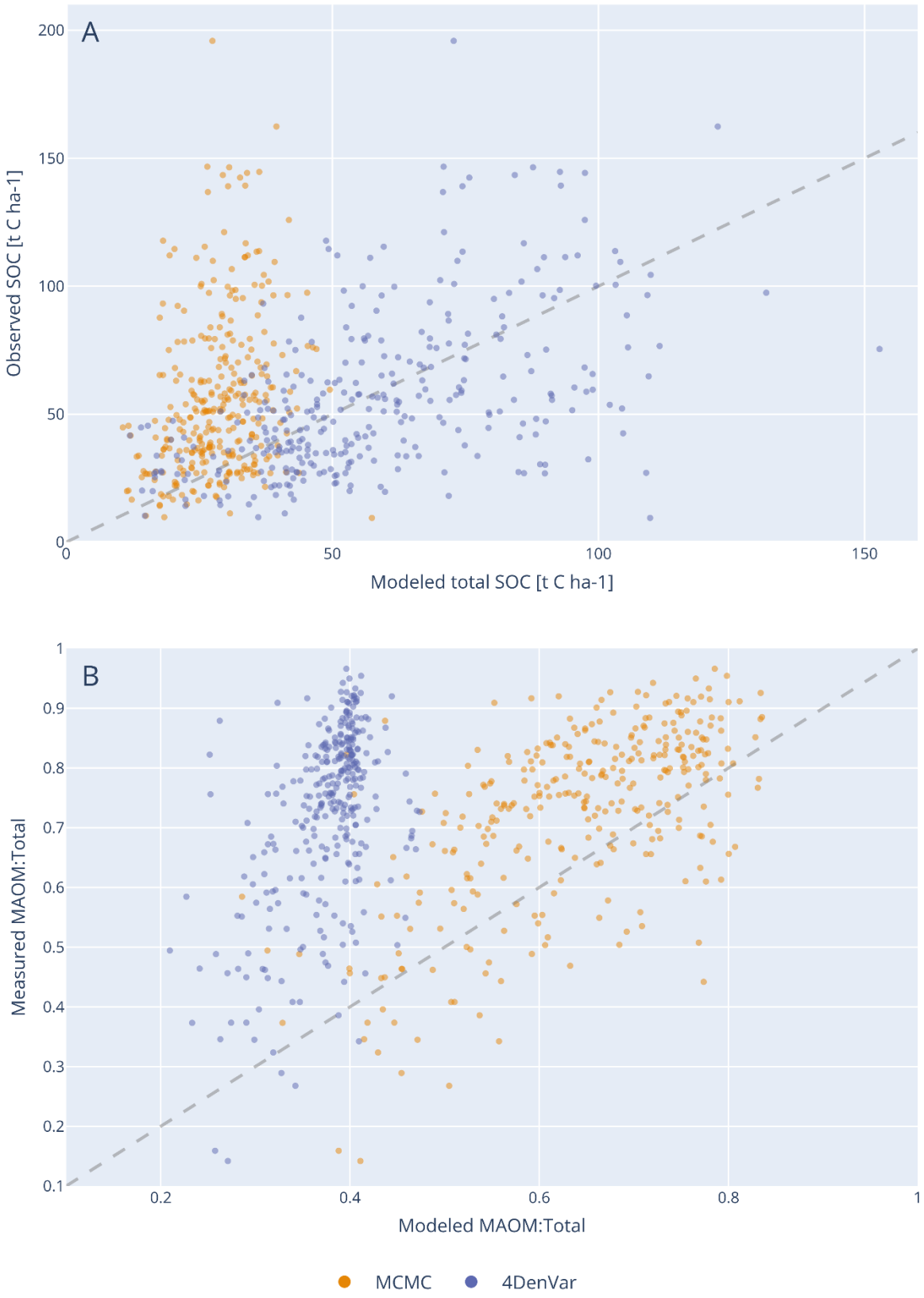
We again changed the expected values to statistically likeliest values. We also removed the baseline parameters as that was an unfortunate remnant of a previous version of the manuscript.

"Figure 4: the MCMC method is not able to simulate the whole range in observed SOC, while both the MCMC method 4DEnVar systematically overestimate the MAOM:SOM ratio (with 4DEnVar not being able to simulate the whole range in measured ratios). As these are calibration results, I would have expected the models to perform better, at least without clear biases. Can a reason be that the ranges in the values of calibration parameters weren't large enough (which is difficult to check by the reader because of the generic parameter names)? Also, please add to the labels on the x-axes that these are the modelled results."

In our view, the two issues are connected. If you look at Figure 4 again, both MCMC and 4DEnVar actually underestimate the measured MAOM ratio, which are very high. The model dynamics struggle to produce both that large of a MAOM fraction while also maintaining the total SOC at the given range, especially with a lower fdoc which would result in a less litter being inputted into the soil.

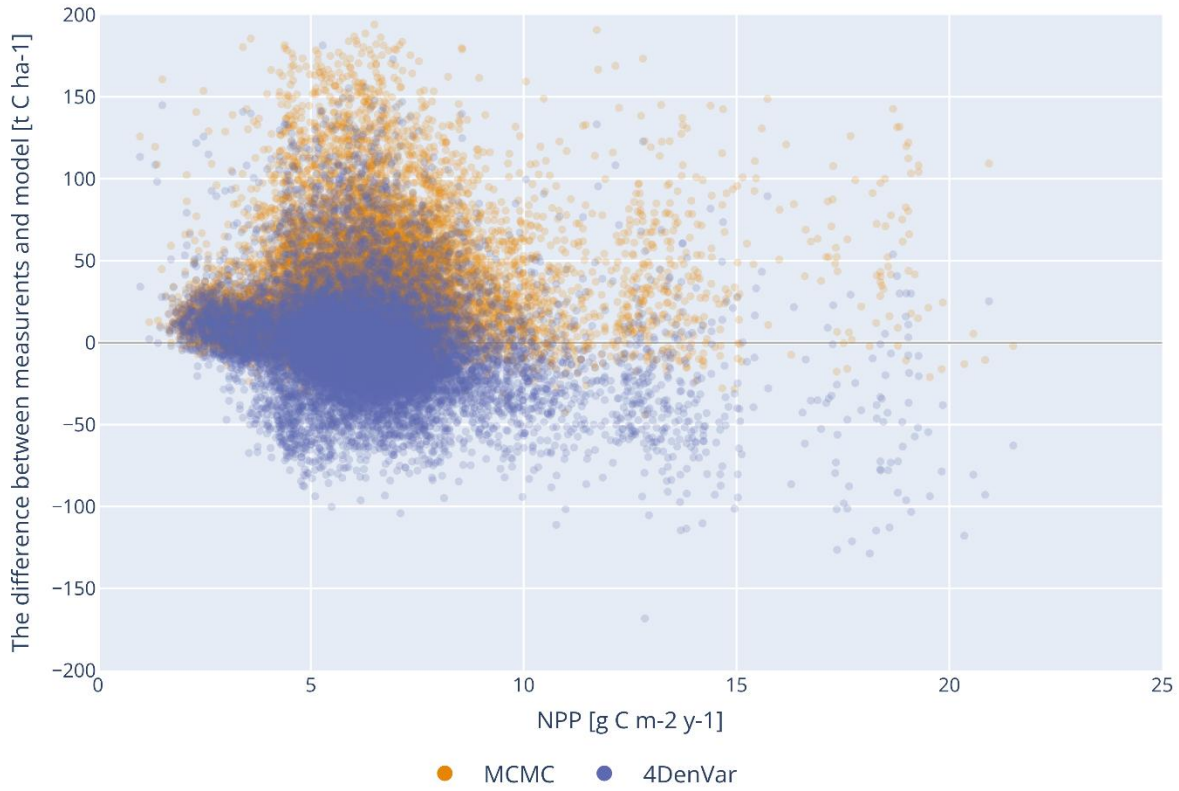
While the parameter value range is a possibility, if you look at the produced distributions in Fig 2, the MCMC decomposition parameter estimates are hitting the lower boundary in a manner where if we lowered those even more, it would lead to those SOC pools essentially not decomposing at all. Thus, to us it is a more of a reflection of the litter issue which we attempted to explain in the discussion.

We have also corrected the figure axis as suggested with the new Fig 4 being:



"Figure 6: please provide a time unit for NPP on the x-axis"

Added to the figure as seen below.



"L 412-425: the interpretation the performance of both methods would benefit from a more quantitative detailed description of the validation results. For example, a scatterplot of modelled versus measured SOM for the validation dataset, combined with different error measures. Currently Fig. 7C is the only figure where the reader can see the model error for the validation dataset."

We have added error statistics to the Results section to provide an easier way to evaluate error analysis of the validation dataset. From line 480:

To examine the impact of the new parameter sets, Figure 3 presents the differences between the measurements and model projections across all the validation sites, while in Table 4 we show both the Root Mean Square Error (RMSE) and mean error (ME) representing bias in regard of the validation dataset for each parameter set. While the 4DEnVar parameter sets produces a somewhat symmetric error distribution around zero in both calibrations, with the higher f_{doc} there is a slight apparent tendency towards positive errors. In contrast, the MCMC error distribution shows a notable lean towards positive errors for the lower f_{doc}, while with the higher f_{doc}, the bias is much reduced. Since the SOC errors here are calculated as the measurement minus the model projection, this means that positive errors reflect the parameter set systematically underestimating the SOC projections. It is notable that with the higher f_{doc}, the RMSE values for the two parameterizations are very closer to each other even with the larger positive bias of the 4DEnVar method.

	f_{doc} 0.15	f_{doc} 0.35

MCMC	42.5 / 27.4	31.3 / 7.4
4DEnVar	29.8 / -1.9	32.0 / 14.2

Table 4: The error statistics for the different parameterizations with regard to the validation dataset. The first value is for the root mean square error (RMSE) and the second for the mean error (ME). The unit for all the values is t C ha^{-1}.

"Figure 7A&B: please add to the labels on the x-axes that these are the modelled results."

Since both panels represent modelled values, we fear that adding that to the x-axis would be misleading about the y-axis values. We are also quite explicit in the figure header that these are modelled values.

"L 444-445: I wouldn't say it's striking that the parameter sets differ from each other, this is an often-observed characteristic of equifinality (see above). I suggest the authors discuss this in more detail."

As discussed more when the subject was first brought up, we are aware of equifinality and that is a core reason why we first conducted the twin experiment as that is the traditional first step in approximating which parameters can be reliably calibrated simultaneously. The statement brought up here was more on how the calibrating method in itself turned out to be such as an element in the equifinality issue.

We have now expanded this paragraph on line 574 to explain both equifinality and our views on our results' contribution to that challenge:

What is striking, though, is how much the parameter sets produced by the two calibration methods in both litter distribution scenarios differ from each, even with the higher f_{doc}, they perform approximately equally well with regard to the total SOC measurements in the validation dataset. As mentioned in Introduction, equifinality, a situation where there exists multiple parameter sets that produce similar model outputs, is a known issue in ecosystem modelling in general and is evidently represented by the results here. The notable element here is that the calibration method itself determines the resulting parameter set as even when repeated, the MCMC calibration approach does not suggest the solution is in the same part of the parameter space as the 4DEnVar results indicate. Generally, twin experiments are efficient first pass to test for equifinality and the challenge can be addressed by reducing the amount of parameters being calibrated, but here there are questions how much those efforts can be relied on in assessing equifinality.

"L 445-446: it's not clear to me how both parameter sets 'perform equally well with the validation dataset', as no error measures for this have been provided, and Fig. 7 shows that there are clear differences between the simulation results for both methods. Therefore, I suggest the authors quantify model performance for the validation dataset, and explain why they interpret the validation results as being equally well between both

methods. In addition, a good test of the effect of the different parameter sets would be to run your validation sites into a predictive mode, using for example an artificial increase in temperature for a couple of decades. If both methods result in a similar change in SOC for each site, you can say they 'perform equally well', but if both parameter sets results in a different change in SOC for each site, you can conclude that the different parameter sets have a different effect when moving away from the steady-state solution."

We have added the basic error statistics to the Results section which hopefully clarifies this part up. Additionally, there was an error on our part as our intent here was to refer to the higher f_{doc} values as with the lower one, the MCMC error has a very notable positive bias.

We have changed the sentence on line 574 to better reflect this:

What is striking, though, is the how much the parameter sets produced by the two calibration methods in both litter distribution scenarios differ from each, even with the higher f_{doc}, they perform approximately equally well with regard to the total SOC measurements in the validation dataset at the given time. It is important to note

Since the calibration dataset is over a large area, the performance is assessed with regards to overall performance for the combined dataset. Thus, the discussion here is not about would a user get a good result for a single site in a specific region with this model, but rather how well they would get result on average when modelling across regions.

It should, though, be noted that our statement within a context about that being the available validation dataset and only applying. At no point in the Discussion section do we suggest that we expect the future projections from these models to be the same under climate change as the parameter are too different. Furthermore, while Fig 7 does show the POM/MAOM values being different between the two calibrations, we do not have those fraction measurements for the validation dataset and thus cannot use that for validation

"L 454-465: it's not clear why these differences in parameter sets are attributed to the measurements (MAOM and total SOC), and not to a potentially inappropriate model structure, parameters values that were fixed incorrectly, or the existence of multiple mimima in the error space. What is the reason for not questioning these aspects of the model calibration process?"

We do think that those other listed components can contribute, especially the multiple minima which is visible in the cost function values we have now added to the Results section, but they are fundamental reasons for equifinality. The reason for our focus on the parameterization method in this work is that the differences appear to be consistent while one would immediately expect the calibration method to be such a component in

the equifinality issue. Besides, we do discuss the impact of the prior parameter ranges in the following paragraphs and bring up missing model processes such a soil moisture.

The model structure, though, is a complicated topic, which is why we have not added an addition into the manuscript. The issue is that, while it might feel natural to assume that a more realistic model structure might avoid this problem, this sort of a model would end up having more parameters for which there is even less prior data to constrain them with.

"L 520-521: this statement is difficult to verify, as neither the error distribution nor the error measures are quantified for the validation (or calibration) results. For example, it is not possible for the reader to asses by which percentage the validation results are off, as only absolute numbers are shown in the plots (for example, Fig. 5 and 7C)."

We have now added the statistics, as requested before. As for the statement, it was meant to be more of a reflection about the general performance considering all the missing processes, not as an objective fact. We have now rewritten the sentence from line 689 to better indicate this viewpoint:

However, when considering the multitude of simplifications made to calculate the steady state approximations using parameters calibrated with data from 322 sites, the error distribution for the 17 000+ validation sites is much narrower than we initially expected.

"L536-537: SOC data alone is indeed often not sufficient to evaluate the performance of SOM models, see Guo et al. (2022; https://doi.org/10.1016/j.soilbio.2022.108780) or Braakhekke et al. (2014; https://doi.org/10.1002/2013JG002420)"

Thank you for the references to the previous and we have now added those to the sentence starting from line 706:

These outcomes emphasise the importance of carefully considering how model performance improvements are assessed with large-scale datasets such as the LUCAS measurement data, since the total SOC seems not sufficient which is in line with previous studies (Braakhekke et al., 2014; Guo et al., 2022).

"L539: assessing the thermal stability of SOM is not the same as a fractionation into POM and MAOM (although they may be related), so the reference by Delahaie et al. is not appropriate as an example of a more efficient fractionation into POM and MAOM."

Thank you for pointing this out. We have changed the Delahaie et al reference to Leuthold et al., 2023. The exact reference is at the end of this response with the other added references.

"L552-554: the conclusion that both methods produce an 'as good validation performance' needs to be supported by a quantitative assessment."

We have now added the statistics to the Results section and discussed them there.

"L556-557: '[...] to notably impact future projections': have such analyses been performed? That would be the ultimate proof to assess how the different parameter sets affected the model performance."

We have already shown that the different parameter sets produce large differences in the model projected POM/MAOM fractions as discussed at the start of this sentence. That in itself does showcase how they would result in different projections as, using an extreme example, if all vegetation would be removed from a plot, there would be a very large difference in how much SOC would be remaining in 10 years if the initial state was produced by the parameterizations presented here.

We do have plans to test the impact of this equifinality in different future projections but considered it outside the scope of this initial paper.

"L 557: I wouldn't call an increase in the portion of NPP going into the soil from 15% to 35 % a 'slight change', as this is more than a doubling"

Fair point and we have adjusted the sentence on line 730 to:

We also conducted a simple experiment to assess the impact of changing how the soil litter input is distributed among different litter pools.

Technical comments

We have made all the corrections listed by the reviewer here.

Added references:

Beylat, S., Raoult, N., Bacour, C., Douglas, N., Quaife, T., Bastrikov, V., Rayner, P.J., and Peylin, P.: Towards the assimilation of atmospheric CO_2 concentration data in a land surface model using adjoint-free variational methods. *Geosci Model Dev*, **18**, 7501-7527, 10.5194/gmd-18-7501-2025, 2025

Braakhekke, M.C., Beer, C., Schrumpf, M., Ekici, A., Ahrens, B., Hoosbeek, M.R., Krujit, B., Kabat, P., and Reichstein, M.: The use of radiocarbon to constrain current and future soil organic matter turnover and transport in a temperate forest. *J Geophys Res Biogeosciences*, **119(3)**, 372-391, 10.1002/2013JG002420, 2014

Brunmayer, A.S., Hagedorn, F., Moreno Duborgel, M., Minich, L.I., and Graven H.D.: Radiocarbon analysis reveals underestimation of soil organic carbon persistence in new-generation soil model. *Geosci Model Dev*, **17**, 5961-5985, 2024

Cambardella, C.A., and Elliot, E.T.: Particulate Soil Organic Matter Changes across a Grassland Cultivation Sequence. *Soil Sci Soc Am J*, **56(3)**, 777-783, 1992

Guo, X., Viscarra Rossel, R.A., Want, G., Xiao, L., Wang, M., Zhang, S., and Luo Z.: Particulate and mineral-associated organic carbon turnover revealed by their long-term dynamics. *Soil Biol Biochem*, **17**3, 108780, 10.1016/j.soilbio.2022.108780, 2022

Leuthold, S.J., Haddix, M.L., Lavallee, J., and Cotrufo, M.F.: Physical fractioning techniques. Encyclopedia of Soils in the Environment, 2, 68-80, 10.1016/B978-0-12-822974-3.00067-7, 2023