Dear Editor,

Below are our thorough attempts at respond to the various valid recommendations and questions raised by the reviewers. We are truly grateful for the level of detailed feedback by them and feel the manuscript is greatly improved by their work.

On behalf of the authors of the manuscript,

Toni Viskari

Contract Agent

JRC-Ispra

Reviewer 1:

“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.

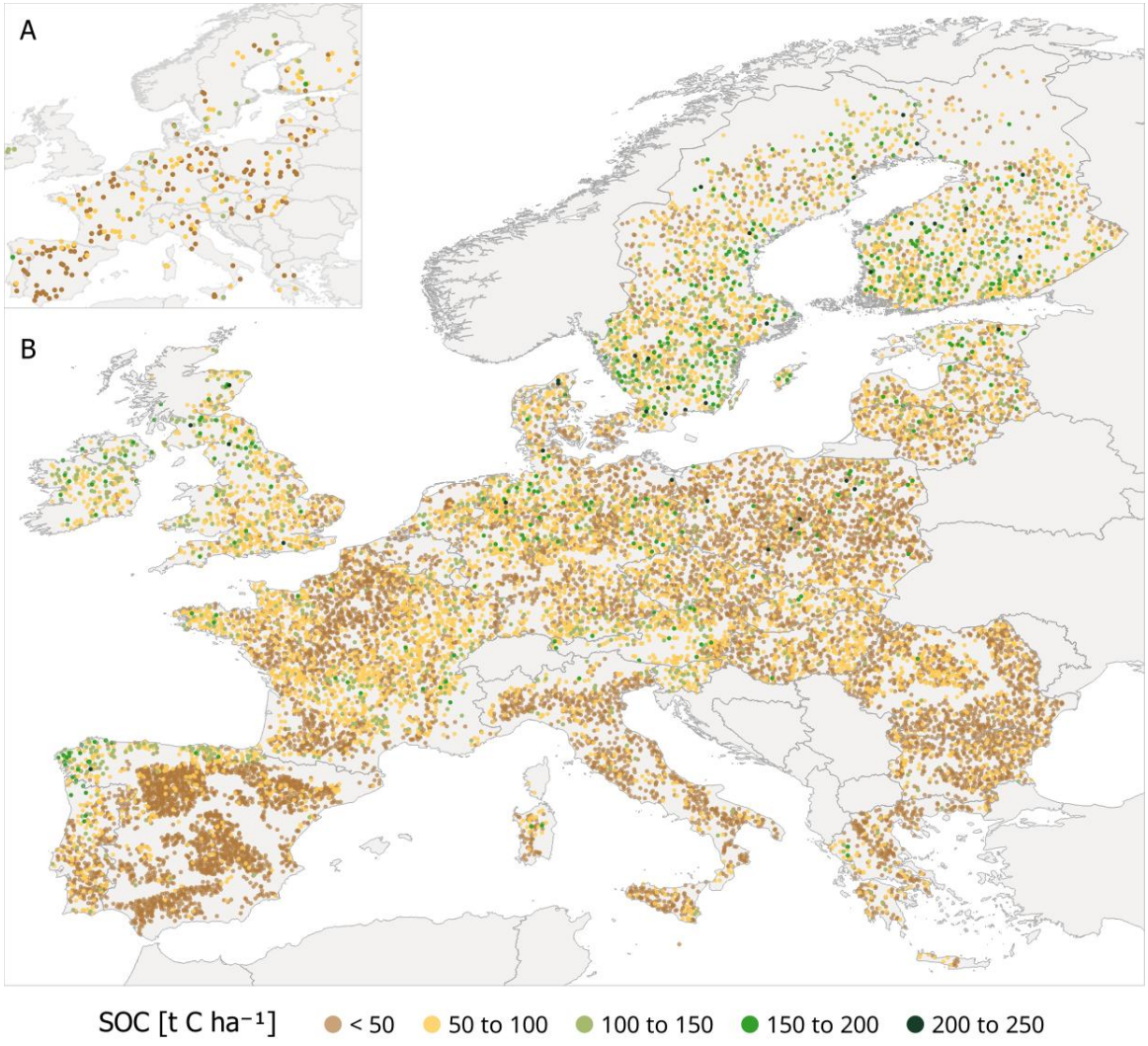
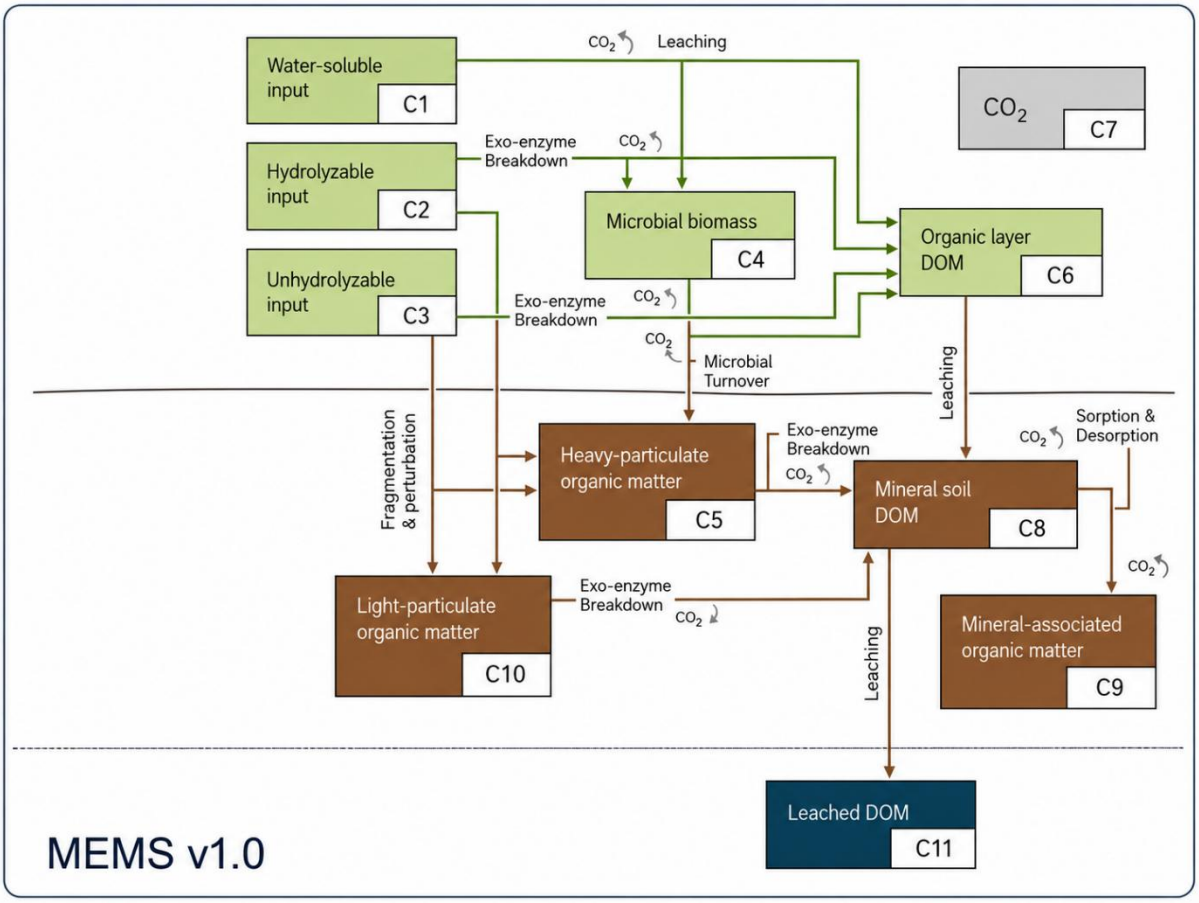


Figure 1: The LUCAS 2009 sampling points across Europe and their SOC stock used for A) calibration and B) validation

"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

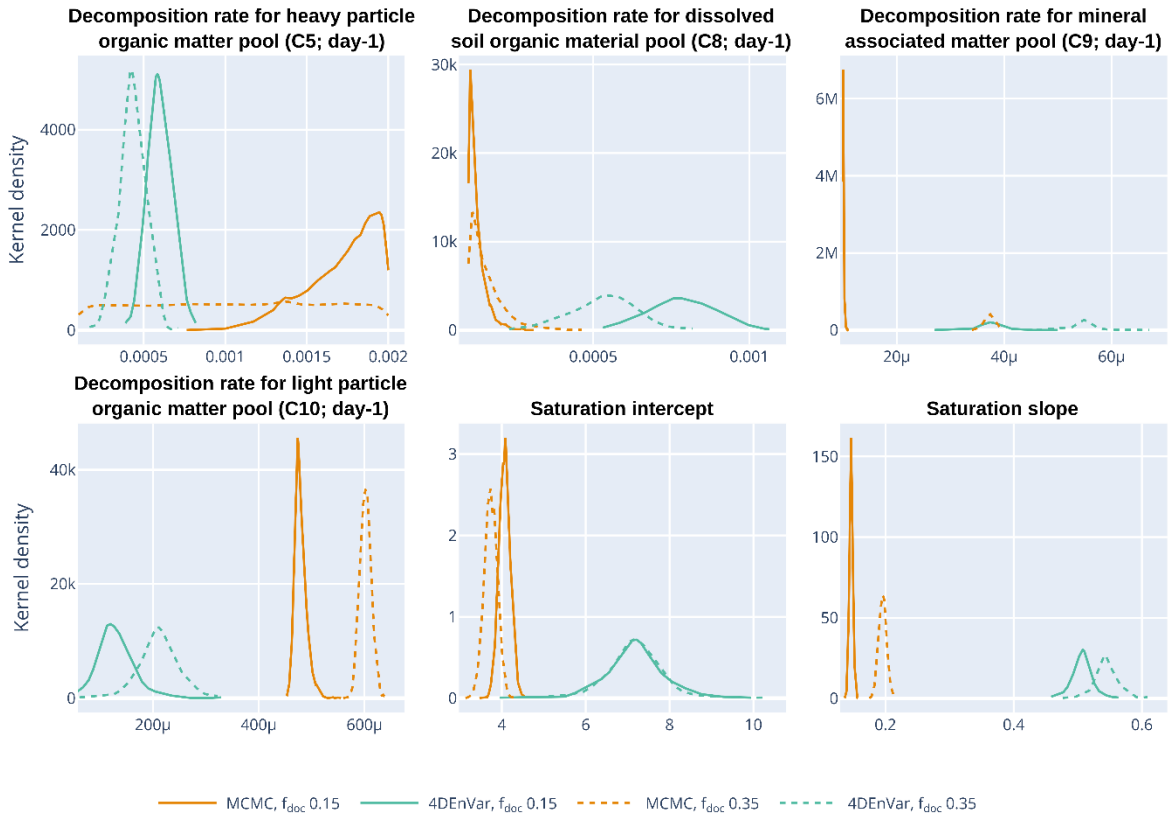
between those two datasets. However, with the synthetic dataset that is not necessary, which allows both methods to just find the same correct parameter set.

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'?"

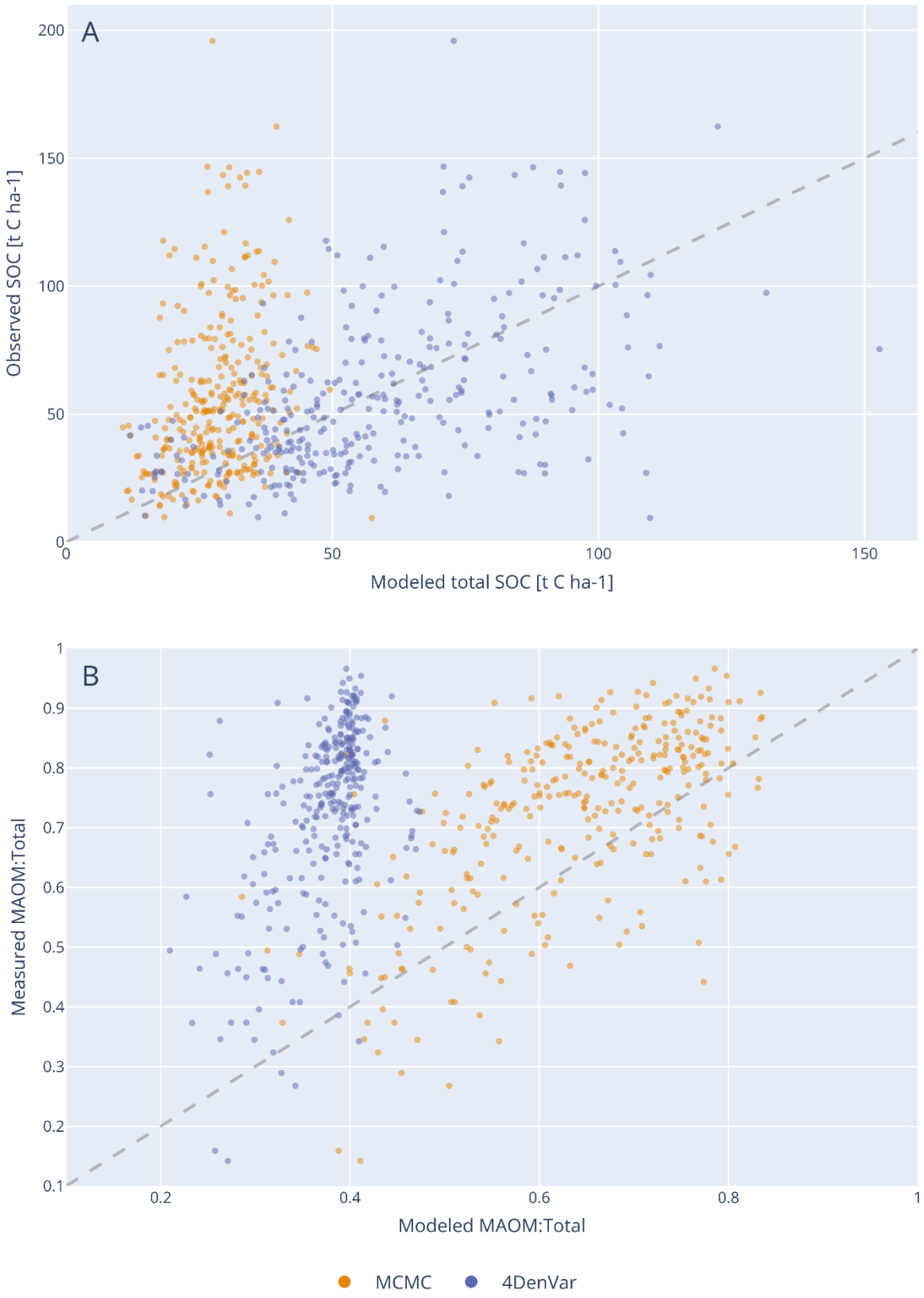
We again changed the expected values to statistically likeliest values. We also removed the baseline parameters as that was an unfortunate remnant of a previous version of the manuscript.

"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.

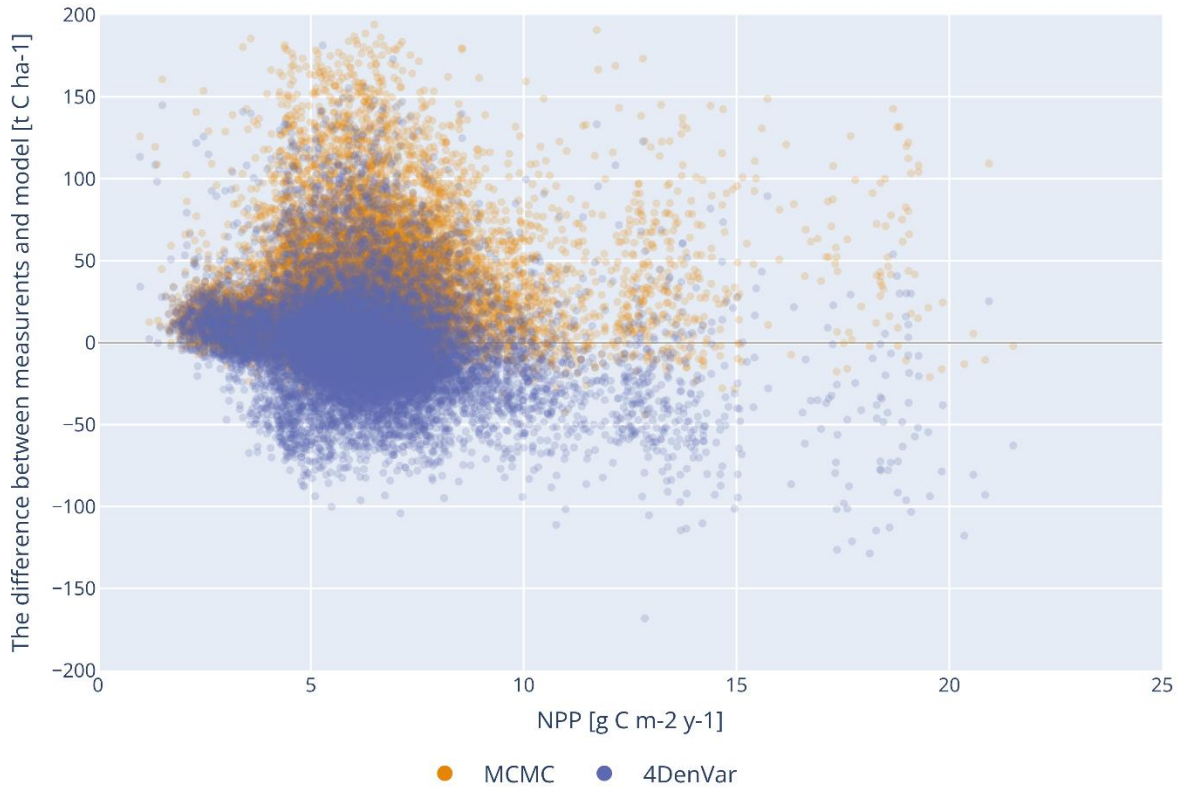
While the parameter value range is a possibility, if you look at the produced distributions in Fig 2, the MCMC decomposition parameter estimates are hitting the lower boundary in a manner where if we lowered those even more, it would lead to those SOC pools essentially not decomposing at all. Thus, to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

Reviewer 2:

“**General Comments**

In their manuscript, Viskari *et al*. tackle one of the most significant bottlenecks in large-scale soil organic carbon (SOC) modelling: the immense computational cost of parameter calibration. By comparing the traditional Markov Chain Monte Carlo (MCMC) algorithm with the novel 4-Dimensional Ensemble Variational (4DEnVar) data assimilation method, the authors provide a highly valuable methodological benchmark. Using the LUCAS 2009 soil inventory, they successfully demonstrate that 4DEnVar can achieve comparable validation performance to MCMC at a fraction of the computational cost, while also revealing how different algorithms navigate conflicting measurement incentives/trade-offs (Total SOC vs. MAOM fraction) and hidden model assumptions (litter input fractions).

The manuscript’s willingness to explore the pitfalls of calibration—specifically equifinality, and boundary-hitting parameter estimations (parameters hitting the ceiling)—makes it an important and refreshing contribution to the field.

However, while the conceptual framework and ultimate findings are strong, the manuscript currently suffers from some mathematical imprecisions/misrepresentations in the Methods section (probably mistakes in equation cross-referencing), as well as a major contradiction between the text and the visual data in the Results section. Moreover, in some cases, key variables are left undefined, mathematical notations in the 4D-Var equations are mismatched (cross-referencing error probably), and the interpretation of the kernel density plots mischaracterises the parameter uncertainty (I will apologise in advance if my understanding of these figures is off). Furthermore, the reliance on *data not shown* for foundational steps of the experiment hinders the reproducibility expected.

In my humble view, addressing the specific comments below will greatly improve the mathematical rigor, visual clarity, and overall readability of the manuscript.

The detailed responses are annotated directly in the attached pdf version of the manuscript.

Citation: https://doi.org/10.5194/egusphere-2025-4999-RC2”

We warmly thank the reviewer for the positive review as well as the many apt recommendations they raise for improving the manuscript.

Line-by-line comments:

As a general comment, the lines mentioned in the responses refer to the track-changes version of the manuscript.

Line 1: “I stumbled quite a bit on the title. Could be something like "Calibration of the MEMS v1 model over a continental soil inventory: a comparison of MCMC and 4DEnVar methods"? Just a polite suggestion”

We have changed the title according to the suggestion

Line 39: “Here, the authors mention several existing SOC models (RothC, MIMICS, Millennial) to establish the current modeling landscape. However, the target model of this study, MEMS v1, is not introduced or contextualised before being abruptly named in the study objectives [L81-]. I recommend adding a brief sentence or two earlier in the introduction explaining what type of model MEMS v1 is, how it compares to the others listed, and why it was specifically chosen for this calibration comparison.”

While this is an excellent recommendation, we chose to elaborate on the MEMS model and why it was chosen later in the Introduction. This part was meant to generally comment on soil carbon model calibration and, along with the recommendation of Reviewer 1, has been altered to better reflect that 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)

To address the specific request here, we have added a brief model explanation starting from line 105:

The model in question simulates organic carbon decomposition separately for above- and below-ground carbon with pathways from surface vegetation matter to the soil pools. In the framework of the MEMS v1, the microbial pool is the central connection between the different SOC states and, crucially, along with the soil properties regulates the amount of carbon stored as long-lived MAOM compounds The SOC pools are for the most part connected by first order dynamics, but the relationship between the microbial and MAOM pool is non-linear. Consequently, there is only a small number of central parameters to calibrate while simultaneously the model steady state cannot be analytically solved, requiring the more costly parameterization process.

Line 95: “Is it MEMS or rather MEMS v1? How would you feel about citing it the first time it is introduced in methods section?”

We removed the mention of MEMS here as this section is not model specific. The following section discusses more on the MEMS model itself and contains already the reference.

As for the MEMS or MEMS v1 point, that is an astute note and another oversight on our parts. For the sake of simplicity, we will retain the abbreviation of MEMS and make this clearer as a choice line 163:

The Microbial Efficiency-Matrix Stabilization V1 (referred to simply as MEMS for simplicity; Robertson et al., 2019) model…

Line 151: “In the opening paragraph of Sect. 2.2, the authors helpfully define the physical meaning of the pols C5, C8, C9 and C10. However, when introducing Eqns. 8-11, pools C1, c2, C3 and C6 are used but are never explicitly defined in the text. For readers unfamiliar with the standard MEMS v1 model architecture, I would add a brief sentence defining what physical components these four surface pools represent.

You might want to define r^eco.”

Excellent point and we have added a description of the other pools to the paragraph from line 172:

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. These 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 since they had been chosen to be representative of 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.

Apologies for us forgetting this definition here and our gratitude for pointing it out. Now the term is explained on line 226:

Finally, the r^{eco} represents the fraction of NPP that is assumed to have been removed from the system due to economic activies (harvest, grazing, etc.)

Line 212: “Immediately following Eq. (13), the text states: "First, it is essentially the same as exponent component in Eq. 8, except that is written it in vector form." Equation 8 calculates litter pool input and contains no exponent. It appears the authors intended to reference Eq. (12). Please confirm and /or correct where necessary.”

This was an error on our part and the equation reference has been changed to Eq 12.

Line 225: “Please confirm and correct equation cross-referencing (Probably meant Eq. 14, right?)”

Another error from our part and has been changed to refer to Eq 14.

Line 340: “The authors state that twin experiments were used to establish that both methods could recover the true parameters from synthetic observations, and that this experiment justified the 4DEnVar ensemble size of 250 members. However, the authors note that these results are "not shown". In the interest of reproducibility and transparency, I strongly recommend that the results of the twin experiments be included, at least in the Supplementary.”

While this is a fair request, it is a bit complicated by what we meant by consistency in this situation. There was no one twin experiment where we progressively increased the size of the ensemble. Rather, with the 4DEnVar, we multiplied twin experiments with each ensemble size. This was because, for example, with an ensemble size of 100 we could have the calibration retrieving the correct parameter set four out of five times, but then it leaves still that fifth instance when it did. Thus, the statement of ‘the ensemble size of 250 performed consistently’, actually means that it still produces the correct twin experiment parameter set through numerous repetitions.

This is unfortunately not something that can be easily conveyed through a figure. But we have attempted to address this by being more explicit when describing the twin experiment and being clearer when bringing the twin experiment results up in the Results section.

The new Twin experiment method section part on line 343:

After having set up the algorithmic framework for both calibration methods for the selected LUCAS data points, the first task was to complete twin experiments. In those, we randomly drew a value for each the parameter chosen for calibration from the uncertainty distributions assigned for them in Table 1. Synthetic observations were generated with the model using the new parameter set. Then, we performed the calibration with both tested methods using these synthetic observations with their associated uncertainties set to be 1 % of those synthetic observations and still using the same prior distribution established in Table 1. This allows us to check if both methods were able to find the correct parameter sets in a situation where the true answer was known. For the 4DEnVar, the additional importance of these tests is to assess the ensemble size dimension required to consistently estimate the correct parameter set. This was accomplished by repeating the twin experiment multiple times with different ensemble sizes and choosing the ensemble size where the calibration always found the correct parameter set. The repetitions were necessary because the 4DEnVar ensemble members are randomly drawn, therefore there are potential situations where a given ensemble size can retrieve the correct parameter set several times in a row, but then fails on the next time.

Then we made the explanation on line 440 more explicit:

For 4DEnVar, the experiments established that an ensemble size of 250 members consistently produced the parameters used to generate the synthetic observations for all repetitions of the twin experiment and, thus, we chose this ensemble size for the 4DEnVar consequents.

Line 347: “You might want to mention this before Fig. 2, as it causes confusion: for example, that Fig. 2 is for the synthetic data.”

Excellent point and we moved the paragraph to be before Fig. 2.

In the text describing Fig. 2, the authors state: "Furthermore, with the higher fdoc value, the uncertainty estimates with both methods end up being narrower..." Visual inspection of Fig. 2 reveals this is incorrect for both calibtaion methods. For MCMC, the dashed peaks (fdoc=0.35) are visibly lower than their solid counterparts, which in a kernel density plot seems to indicate greater uncertainty. Of course, hitting a ceiling at lower fdocs might be th reason, if I get the logic right. For 4DEnVar, the uncertainty distributions do not appear to narrow at all: rather, the dashed curves mostly simply translate (shift along the x-axis) while maintaining virtually the exact same shape and width as the solid curves (except sclcept). This rigid shape is likely an artifact of the strict prior uncertainty (10% standard deviation) the authors had to impose. If my reasoning holds. please revise the text to accurately reflect the visual data. Critically, for MCMC calibration, the higher fdoc shifted the parameters into more realistic spaces, but it did not result in narrower uncertainty estimates for either method.”

Line 358: Assuming I understand Fig. 2, here is my comment:

The reviewer is completely correct with this analysis and I, as the primary author, am not certain what made me make this erroneous statement. We have corrected the sentence on line 447:

Furthermore, with the higher f_{doc} value, the parameter distributions produced by 4DEnVar remain approximately as wide even when they shift. Meanwhile with the MCMC calibration it produces wider distributions which represents larger uncertainties.

Line 370: “Grammatically confusing?”

We made the sentence clearer starting from line 484 :

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 near zero.

Line 385: “You might want to show this, at least in the Supplementary“

Fair request and we have added it as the Supplementary figure 2 as well as a reference to it.

Line 389: “The caption for Fig. 4 lacks critical parameter context. While the main text clarifies that these plots represent the lower fdoc scenario (0.15) , the caption itself omits this. Because figures must be self-contained, please add [fdoc=0.15] somewhere in the caption for Fig. 4 caption.

Furthermore [refer to L385- comment], the text mentions that the calibration fits for the higher fdoc (0.35) were also examined but are NOT SHOWN. Because the fdoc comparison is a central component of this paper's narrative, please provide the corresponding plots for the fdoc=0.35 calibration fit in the Supplementary so readers can visually verify how the higher litter input improved the model-measurement agreement or not.”

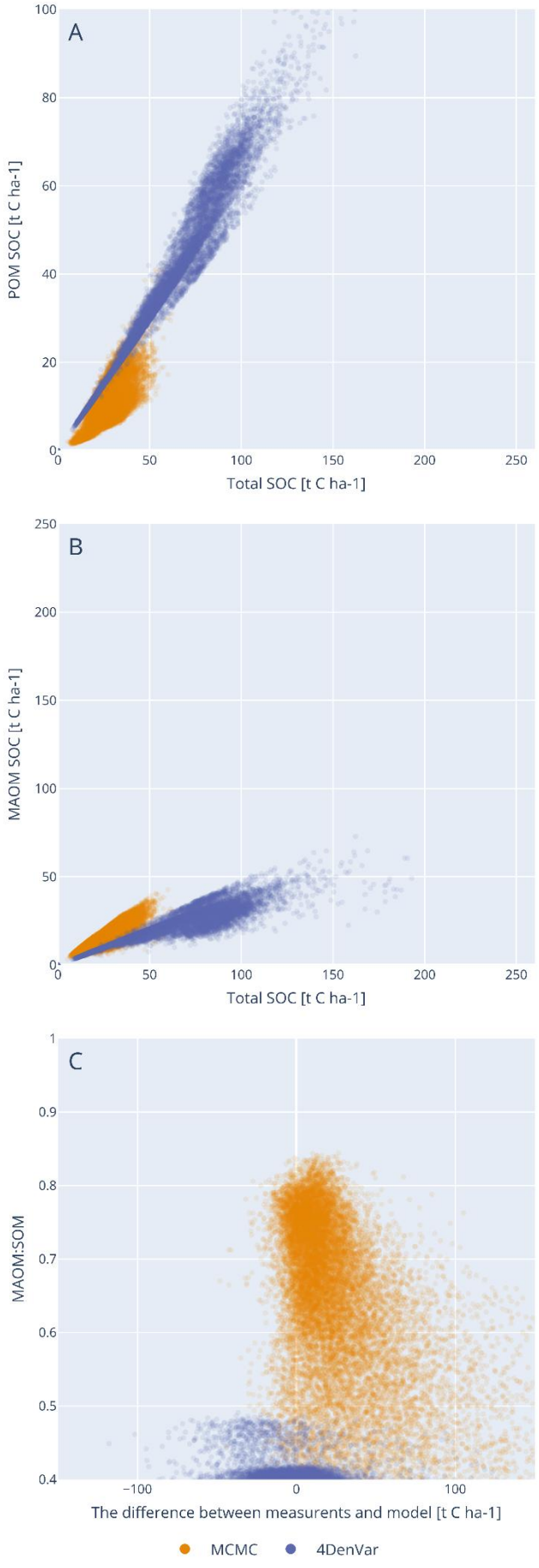
We have added the fdoc clarification to the caption, our gratitude for you noting that, and, as mentioned before, have also included the higher fdoc results in the Supplementary material.

Line 407: “Again, I would show this i the Supplementary or say nothing of it.”

We removed the reference to the ecosystem specific behaviours, upon rereading, as a topic that would be deserve a deep dive than a casual mention like this.

Line 424: “Again, it might be helpful to show this in the Supplementary”

Agreed and added the Supplementary figure 3 as seen below:


A
POM SOC [t C ha-1]
Total SOC [t C ha-1]
B
MAOM SOC [t C ha-1]
Total SOC [t C ha-1]
C
MAOM:SOM
The difference between measurents and model [t C ha-1]
MCMC
4DenVar

Supplemental Figure 3: The model projected a) POM, b) MAOM stocks in relation to the total modelled SOC stocks as well as c) The MAOM:SOM ratio in relation to the model error across the LUCAS sites when f_{doc} is 0.15.

Line 441: “Local minima?”

We changed the “first stable parameter set” to “local cost function minima”

Line 454: “This is interesting. I loved reading this.”

Thank you for this compliment.

Reviewer 3:

"This manuscript by Viskari et al. presents the comparison between the established MCMC and emergent 4DEnVar calibration approaches for the MEMS v1 soil organic model using the continental-scale soil inventory dataset. The authors calibrate the model using a subset of sites where both total SOC and physical fractions were available, and validate these calibration across over 17,000 sites at a continental scale. Then they include a parameter experiment (changed from 0.15 to 0.35) to test the effects of litter input assumption on resulting parameters sets and model projections.

The authors find that MCMC and 4DEnVar calibration methods could achieve comparable performance in terms of total SOC validation with different parameter sets and different internal representation of the SOC state, while the later method is far more efficient. Changes in litter input assumptions affect both calibration methods but lead to general similar behaviors. Thus, 4DEnVar would be a potential tool for efficient calibration, yet more concrete evaluation approach for SOC modelling is needed.

This manuscript makes valuable contributions to the core issue of equifinality in the field of modelling by linking structural discrepancy between calibrating against total SOC and SOC fractions. Comparison between two methods is a practical finding for any modeler considering adopting these method and datasets.

The manuscript is generally well written. The general comments and specific comments are as following:"

We appreciate the positive feedback and the very helpful recommendations.

Line-by-line comments:

As a general comment,

"Line 9-19: It would be beneficial to mention the second object and its findings of this manuscript, i.e. how NPP litter assumptions affecting the calibration and projections, to ensure that the abstract reflects the full scope of this research."

Thank you for highlighting our error in not mentioning the NPP experiment in the abstract. We have now added this to abstract while condensing the text in order keep the character count down. Starting from line 12, the abstract now reads:

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.

"Line 18: as well -> as well as"

Corrected

"Line 69-80: While motivation for this study lies in the computation challenge of MCMC method, the introduction of 4DEnVar method would be expected to be introduced in a way that emphasizing on its relative computation cost with concrete references or statistics to strengthen the motivation of this work."

Excellent point and we have now reworked the paragraph to stress the computational benefit from the start. However, we were not certain what sort of concrete reference/statistics would be considered appropriate, here, as this is among the first studies, to our knowledge, that has compared the two calibration methods in this manner and even then, the benefits would be expected to be very system specific.

In order to address this request, we have added a reference to Beylat et al. (2025), where they compare the performance of a land surface model when calibrated with 4DEnVar calibration and with the traditional 4-Dimensional variational assimilation in the context of atmospheric CO2 concentrations. Hopefully this will be a sufficient concrete a reference to address this suggestion. The paragraph is now, 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, uses the ensemble to sidestep this requirement by simultaneously running multiple simulations with different parameter sets instead of an iterative solution. The 4DEnVar method uses the ensemble to sidestep this requirement by simultaneously running multiple simulations with different parameter sets instead of an iterative solution. While to our knowledge there haven't been previous studies within the ecosystem modelling analysing

the performance of the 4DEnVar to that of MCMC, in Beylat et al. (2025) the 4DEnVar method is compared to the original 4DVAR method in a very specific synthetic experiment. Within that scope the 4DEnVar was shown to be more effective than the original version, but it is only the first step in evaluation.

"Line 81-82: It is unclear for the reader why choose MEMS v1 model for this study, as this model is mentioned only at line 81. Would be nice to introduce this model already after mentioning the need to separate different SOC fractions."

We have added a brief model description as well as a justification for the choice starting from line 105:

The model in question simulates organic carbon decomposition separately for above- and below-ground carbon with pathways from surface vegetation matter to the soil pools. In the framework of the MEMS v1, the microbial pool is the central connection between the different SOC states and, crucially, along with the soil properties regulates the amount of carbon stored as long-lived MAOM compounds. For our purposes here, the model is ideal as for the most part the SOC pools are connected by first order dynamics, but the relationship between the microbial and MAOM pool is non-linear. Consequently, there is only a small number of central parameters to calibrate while simultaneously the model steady state cannot be analytically solved, requiring the more costly parameterization process.

"Line 82-84: The advantage of using the LUCAS dataset can be briefly summarized. As a side focus of this paper, I would also like to know the challenges and limitations of using large-scale datasets for model development."

We have expanded this part with an explanation the benefits of using the LUCAS dataset starting from line 115.

Because this LUCAS dataset contains measurements from thousands of plots across Europe and, thus, represents many different types of ecosystems as well as climate conditions, it allows to test a wider performance of the model calibration. One of the advantages was the level of standardisation in sample collection and analysis, the latter done by a unique laboratory, Furthermore, for a small subset of the chosen LUCAS dataset, the POM/MAOM fractioning also had been done, which provided more nuanced information for the calibration process. While Lucas is a standardised framework for SOC, was not specifically designed to assess the MAOM stocks.

As for the challenges and limitations, we have added a small paragraph into the end of section 2.1 about those relating to the dataset here. Now, from line 155, there is the following description:

While the benefit of the LUCAS dataset is its large spatial representation and inclusion of measurements from multiple different ecosystems, the execution of such a vast measurement campaign introduces different source of errors from sampling, labelling, analysis etc. Thus, it is almost more apt to be considered as a combination of several independent campaigns done with the same protocols, instead of a single consistently controlled campaign. Additionally, although locations of the measurement are known, we have the make the assumptions that the available driver data are representative for the actual conditions at the measurement plot.

"Line 85-86: From the previous context, I can't see why can draw the hypothesis of 'fit to the same degree'."

We rephrased the hypothesis according to the feedback and now it reads from line 122:

Our hypothesis is that the 4DEnVar improves the model fit to a sufficient degree that, along with the reduced computational cost, it can be considered as valid calibration approach for SOC models as the MCMC

"Line 87-91: The two objectives of this study are not articulately explained. Why this 'simple experiment' important in the context of 'comparison of calibration methods', and how is the first and two objectives inform each other?"

We have made the explanation more explicit starting from line 124:

Specifically, there are two objectives for the work presented here: the first is to test if the much faster 4DEnVar calibration performs as well as the MCMC calibration and examine if there are any meaningful differences in the resulting parameter sets; the second is to conduct a simple experiment where we made a change on how the NPP litter input was partitioned. The reasoning for the latter objective is that one of the core benefits of the faster calibration method is that it allows testing how different assumptions impact the parameterizations. Because of this, if there are differences between the results of the two calibration methods, it is important to assess if the general behaviour of the parameterizations remains the same even under different assumptions.

"Line 95-102: I wonder how presentative the subset of 350 samples is, which might be a big source of validation error. Would you please show the distribution of this subset in Figure 1 and provide more information regarding the contained context."

The subset in question was randomly drawn with the distribution of ecosystems kept constant. Thus, it should be as representative of the whole as possible. To make this clearer, from line 140, we have now added this detail:

The latter were randomly drawn from the all the measurements with the only constraint being that both datasets were similarly distributed across ecosystems with approximately 73 % being grass- or croplands with the rest being various forest types. The representativeness of the chosen 350 measurements points is elaborated upon in Lugato et al. (2021).

We have also added a panel to Figure 1 that shows the spatial distribution of the calibration datapoints. The Figure 1 is now:

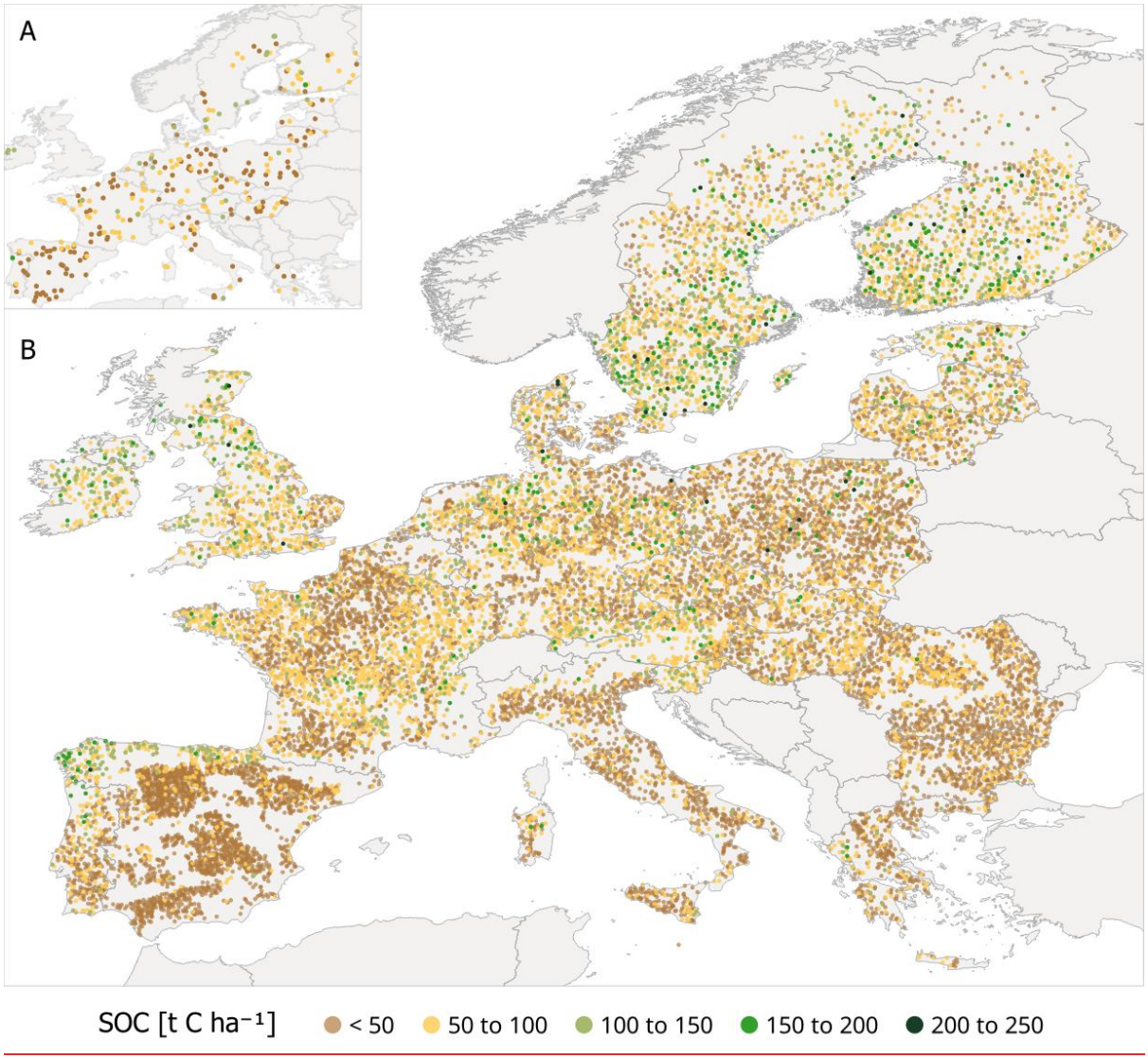


Figure 1: The LUCAS 2009 sampling points across Europe and their SOC stock used for A) calibration and B) validation.

"Line 105-109: I wonder what is the intention to include agricultural soils into simulation, where the-steady state assumption might not hold true?"

We simply chose to include all the representative ecosystems in Europe into the calibration. Especially since even when using SOC models in agriculture, most of them have also been calibrated with steady state assumptions since we very rarely have a reliable detailed initial state that could then be used as a starting point when calibrating time series data. Although in this particular environment, Europe has been similarly cultivated for so long that the steady state assumption is more reliable here than in many other regions.

Furthermore, even in grasslands and non-commercial forests, the steady state assumption is fraught now due to climate change that is affecting all soils in the world.

Thus, there is no such thing as a steady state, but it is unfortunately assumption we still have to rely on.

This is still a valid comment, though, and we have added a line to highlight this as an additional uncertainty source in section 2.5 as we explain the steady state approach itself there. The addition from line 365 is:

It should be noted that with agricultural soils and commercial forests are expected to have a large variability in litter input over a given time window, which does raise challenges for the steady state approach. We are still including those data points in the analysis here as this is intended as a general calibration across European ecosystems and there is no additional data to constrain those specific ecosystems, but this is expected to be an additional uncertainty source.

"Function 8-11: What is r^eco?"

Apologies for us forgetting this definition here and our gratitude for pointing it out. Now the term is explained on line 226.

Finally, the r^{eco} represents the fraction of NPP that is assumed to have been removed from the system due to economic activies (harvest, grazing, etc.)

"Line 162: Wrong equations reference?"

Yes, corrected to 8 to 11.

"Line 212: Wrong equation reference?"

Equation reference corrected to 12.

"Line 224: Wrong equation reference?"

Equation reference corrected to 14.

As the primary, my sincere apologies for the wrong equation references here. There was a large restructuring of the manuscript at a late stage of preparation, and these were unfortunate remnants from that. Our gratitude for reading with such attention to notice these errors.

"Line 268-270: Hard for me to follow the description of the twin experiment. What are the baseline parameters for the parameters you calibrate in this model? How do you generate the perturbed parameter set?"

We expanded the paragraph on line 343 in order to clarify the raised points:

After having set up the algorithmic framework for both calibration methods for the selected LUCAS data points, the first task was to complete twin experiments. In those, we randomly drew a value for each the parameter chosen for calibration from the uncertainty distributions assigned for them in Table 1. Synthetic observations were generated with the model using the new parameter set. Then, we performed the calibration with both tested methods using these synthetic observations with their associated uncertainties set to be 1 % of those synthetic observations and still using the same prior distribution established in Table 1. This allows us to check if both methods were able to find the correct parameter sets in a situation where the true answer was known. For the 4dEnVar, the additional importance of these tests is to assess the ensemble size dimension required to consistently estimate the correct parameter set. This was accomplished by repeating the twin experiment multiple times with different ensemble sizes and choosing the ensemble size where the calibration always found the correct parameter set. The repetitions were necessary as, because the 4DEnVar ensemble members are randomly drawn, there are potential situations where with a given ensemble size it retrieves the correct parameter set four times in a row, but then fails on the fifth time.

"Line 299-304: Please provide the full name of the MODIS product. And not clear for me that why the NPP dynamics are not expected to meaningfully affect the modelling results."

We added the product name to the paragraph starting from line 397. Additionally, the reason we assumed that the NPP assumption wouldn't impact the results was the total annual NPP input stays the same. That is now made explicit in the text.

For Net Primary Production (NPP), first the average annual NPP over the decade 2000-2010 is extracted from the MODIS product MOD17A3 (Running et al., 2004) grid cell overlaying each LUCAS point. Then, a standard sine function is used to distribute the NPP across the year in order to produce the daily litter input. This approach was used instead of an averaged MODIS NPP annual time series as the NPP reflects the time when the atmospheric carbon is allocated into vegetation, not when the vegetation becomes litter input. Hence, we simplified the time series here and, since the total annual NPP remains the same, it is not expected to affect the modelling results to a notable degree.

"Line 363: I think the baseline parameters are not presented here? Also I wonder if you could provide uncertainty measures in this table? 0.3s5 -> 0.35."

Thank you for pointing out the baseline parameter line, a remnant of an older version of the manuscript. We have also corrected the 0.35 typo. The table header on line 478 is now:

Table 3: The expected parameter values produced by the different calibration methods. The first value is for f_{doc} 0.15, the second for f_{doc} 0.35.

As for the uncertainty measures, we have already presented those in the Figure 2 with the parameter distributions. We have also now added the standard deviations in Supplemental Table 2. Originally, we tried to include them in table 3, but in our opinion the resulting table was far too cluttered.

"Line 376-377: I wonder if you could provide qualitative statistics for the error distribution. As the difference between two methods with f_doc=0.35 can't be intuitively recognized."

Another oversight on our part and we are grateful for this review as well the others to drawing attention to it. We have now added the error statistics to the manuscript and expanded the Results section to discuss these results. 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}.

"Line 401-403: The spatial pattern in Figure 5 seems to be coincided with the contexts in Line 404-408, and conclusions in 422-425. I wonder if you could discuss a bit more the regional patterns and related driver data limitation, model structure limitation, impacts in model performance evaluation in the discussion section. To make it more concreate in depicting the challenge in modelling with large-scale dataset."

An excellent suggestion that adds depth to our results in this work. We have expanded the discussion on two parts on the regional impacts.

First, when discussing the role of limited soil moisture dynamics hinder the model in question, we added a sentence on line 614 to connect those to the biases we see in the Iberian peninsula:

This could be a partial explanation for the Iberian peninsula error biases visible in Fig 5 as the soil moisture dynamics are much more complicated in arid climates vulnerable to drought (Almendra-Martin et al., 2021)

Then, in the NPP assumption section of the Discussion, we wrote a new paragraph starting from line 672 where we touch on the spatial element of that issue:

Adding to the challenges discussed above is that the various assumptions are not expected to be spatially homogeneous even in the same ecosystem type. For instance, the Nordic countries, especially Sweden and Finland, are dominated by economic forests where the NPP-to-litter pathway is heavily impacted by the growth stage as newly growing forest will have a large NPP, but not a corresponding amount of litter due to mortality. This could be connected to bias seen in the northern Europe in Fig 5. Another example would be agricultural ecosystems as climate conditions affect which crops will be dominant in a given region. The type of crops naturally affects its traits as, for instance, the root depth distribution, which in turn is expected to impact the soil carbon stocks (Fan et al., 2016). These various components could be a reason why when analysing global soil databases, there is a weak statistical relationship between NPP and SOC despite that dynamic being well understood (Luo et al., 2021).

"Line 444-453: looks more like results for me, and these are not brought up in the results section."

We agree and have moved the comments about the cost function to the Results section to line 508.

As for this paragraph here, based on comments from another reviewer, we have expanded it to discuss in more depth equifinality and how the results here contribute to that challenge. From line 574:

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 at the given time. Equifinality, a situation where there are multiple parameter sets that produce similar model outputs, is a known issue in ecosystem modelling in general (Sierra et al., 2015; Marschmann et al., 2019), but the surprising element here is that the calibration method itself determines the resulting parameter set. 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.

"Line 466-492: I wonder if you could provide a bit more information on the potential solutions to solve the prior impacts on calibration. So that this part would be more useful for practitioners considering using 4DEnVar model for similar applications."

This is a really good request and something we should have addressed in the original version of the manuscript, even if there are naturally no easy answers.

In response, we have brought up the importance of providing uncertainties with the measurement sets as well as putting more effort in producing the parameter prior distributions as there is often a tendency to just give uniform distributions. We also point out that when using multiple datasets for calibration, it is important to first confirm if they are compatible with each other.

Finally, while we have only done a single calibration cycle with the 4DEnVar in the study here, the results from a single calibration can technically be used as the prior range for the next one. The benefit of doing this is that by repeating the process multiple times until the results themselves do not change is that it allows addressing the concern that the initial prior range is too far away from the actual value.

To provide a hypothetical example of this, let's say we have a parameter where the optimal value is 2, but because our prior is too badly defined, our expected value is set 6. With one calibration cycle, the estimated value can end up around 4 just due to the sampling of the prior distribution. Yet by repeating the calibration in this situation, the next posterior distribution would be nearer to the correct value.

Why we have not done this, though, and why it is not as straight-forward as it might seem is that this repeating cycle reduces the impact of the prior range which goes against Bayesian philosophy. Furthermore, with each iteration cycle the uncertainty range is reduced which, in turn, results in unrealistically high trust in the results. Thus, while the repeated calibration cycles is a possibility, it needs to always be implemented with

Additionally, there is also a practical solution in having the 4DEnVar algorithm to sample outside the prior distribution that would help with some of the issues, but since that is more of a question of the parties creating the algorithms than the user, we do not mention that here.

The new paragraph on line 634 reads:

Naturally this underlines the overall importance of providing reliable measurement uncertainties along with measurements themselves, but that is not something a model user can simply produce by themselves. When implementing the calibration, based on the results here we would recommend of initially looking through the calibration data and confirming that all the values are sensible for the model/system being calibrated. As a more practical solution, it is possible to repeat the 4DEnVar calibration multiple times by using the previous posterior distributions as the priors to the next cycle. This way it is possible to ensure that the resulting parameter set is not simply because the prior had been set too far from the correct value and thus partially reduce the impact of the assigned prior distribution. However, the downside of repeating the calibration cycle in this manner is that not only does it reduce the impact of the prior, but each iteration reduces the resulting uncertainty distribution. Thus, the final parameter distributions would be artificially too confident. While the repeated calibration is a worthwhile tool in certain circumstances, it always needs to be implemented with great care and consideration.

"Line 500-511: Current discussion mainly focuses on the challenges we are facing right now. It would be beneficial to put it in a broaden ecological context, and discuss about the ecological outcomes of changing the f_doc parameter. This might help to assess whether the chosen f_doc parameter is ecological reasonable."

Reasonable request, although it is important to note that within the context of the model implementation here, the f_{doc} value actually govern the litter amount directly deposited into the soil. Thus, it reflects more what is the division between the above- and below ground biomass distribution for different plant species and how we reflect that in our modelling work affects many different facets of the results.

We have expanded this paragraph, and splitted into two, to discuss the ecological representation of that variable. Now, from line 654, the new paragraph reads:

What complicates future work is that coefficients associated with litter input are challenging to calibrate simultaneously with parameters associated with SOC decomposition, as their influence on the SOC overlap too much. It is important to note that while the focus in this experimentation has been the f_{doc} value, what it actually represents is the assumption of dividing NPP between upper- and below ground biomass as it reflects the amount of litter deposited directly into the soil. This is a central assumption that has to be included in some manner in SOC modelling and is represented by the plant species traits assigned to the surface vegetation. This highlights why better understanding of the vegetation qualities of the ecosystem being modelled is important for calibrating even simple SOC models.

As for even attempting to calibrate the NPP/litter coefficients simultaneously would first necessitate determining which exact coefficients would be calibrated. For example, in our case, there is first the question how well the MODIS NPP product represents reality for different systems. Then, part of that NPP is removed to represent economic activity before it is distributed to the four MEMS initial pools based on the three coefficients. Any of these three parts can be altered to change the final NPP input to the soil in different ways, but there is really no certainty at the moment what is the correct manner to better regulate the NPP based litter input. This complicated relationship in the surface vegetation driving litterfall and the SOC state has been shown in prior work such as in Raczka et al. (2021). There when they used remote sensing data to constrain their model state, while this improved their modelled aboveground biomass and carbon exchange accuracy, it also caused their modelled SOC accuracy to decrease because they were only using the aboveground data for both systems.

"Line 552-554: I wonder if this sentence should be stated more carefully, by specify the NPP assumption, while MCMC resulted in lower J under certain circumstance."

Good point, although this statement was in regard of the validation dataset, thus we did not weight this from the J perspective. We have now rewritten the sentence from line 724 onwards as:

In our work presented in this article, we have shown that 4DEnVar parameterization produces the approximately same RMSE for the validation dataset as the traditional and more cumbersome MCMC DEzs algorithm when the soil litter input is increased and actually outperforms in this metric the MCMC with the lower litter input.

"Line 557-559: The conclusion of the second object of this work is rather vague, and didn't explicitly summarize the ecological meaning. Otherwise, it looks like just a repetition of the first object's finding, without drawing importance of the second object."

More than a valid criticism and we have expanded the conclusions here to both be more concrete as well as stress how this reflects the role of ecosystem related assumptions. On line 730:

We also conducted a simple experiment to assess the impact of changing how the soil litter input is distributed among different litter pools. These results showed that while the litter input adjustment did impact the calibration, the general model behaviour produced by the two calibration methods remained similar. This implies, if it holds true with further testing, that the differences between the behaviours of the two calibration methods are not dependent on the driver data. Another facet of these results is that it confirms how large of an impact ecosystem related assumptions have on the resulting calibrations.

"Line 562: Might be nice to add one sentence at the end and return to the broaden context of SOC modelling goals."

We considered this a good suggestion and added the following starting from line 740:

This will make it more pragmatically possible to assess how various assumptions impact ecosystem model results as well as better include those uncertainties in future projections as the various drivers are altered by climate change.

Added references:

Almendra-Martin, L., Martinez-Fernandez, J., Gonzalez-Zamora, A., Benito-Verdugo, and Herrero-Jimenez, C.M.: Agricultural Drought Trends on the Iberian Peninsula: An Analysis Using Modeled and Reanalysis Soil Moisture Products. Atmosphere, 12(2), 236, 10.3390/atmos12020236, 2021

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

Fan, J., McConkey, B., Wang, H., and Janzen, H.: Root distribution by depth for temperate agricultural crops. *Field Crops Res*, **189**, 68-74, 10.1016/j.fcr.2016.02.013, 2016

Luo, Z., Viscarra-Rossel, R.A., and Qian, T.: Similar importance of edaphic and climate factors for controlling soil organic carbon stocks of the world. *Biogeosciences*, **18(6)**, 10.5194/bg-18-2063-2021, 2021