“**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.

Line 350: “Assuming I understand Fig. 2, here is my comment:

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.”

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."

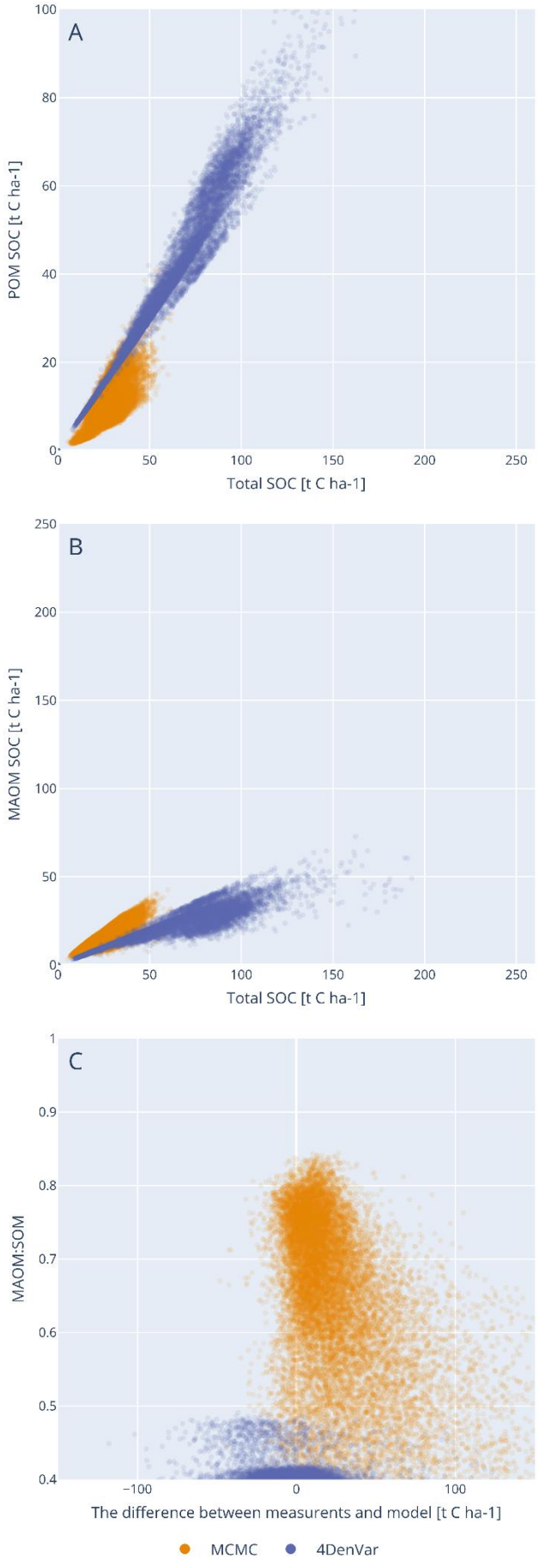
We have added the fdoc clarification to the caption, our gratitude for you noting that, and, as mentioned before, have also included the higher fdoc results in the Supplementary material.

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.