In the document, feedback from the reviewers is written in italic face, while our responses are written in green. Screenshots of parts of the manuscript that have been changed are shown in boxes.

Reviewer 2

This manuscript addresses a fundamental and timely issue in soil organic matter (SOM) modelling: the mismatch between model complexity and data availability, which leads to parameter non-identifiability, equifinality, and ultimately large uncertainty in model predictions under perturbations. The study is conceptually important and provides a clear demonstration that models calibrated to steady-state observations can produce substantially divergent predictions when subjected to changes in carbon inputs. This has important implications for the reliability of current ecosystem and Earth system models.

We thank the reviewer very much for taking the time to read our manuscript and for providing detailed and very constructive feedback. This, together with your indications of parts of the manuscript that weren't properly described, significantly improved the quality of the manuscript. This is greatly appreciated. Please find our responses to the feedback below.

Below I provide specific comments.

Line 6-8: The statement that the number of identifiable parameters "remained limited to five even under the most data-rich conditions" may be somewhat overstated. The "most data-rich" scenario considered here still represents a relatively limited set of observations, and the result is likely dependent on the specific model structure and data types used. This limitation should be more clearly acknowledged.

Thanks for pointing this out, these specific results are indeed dependent on the model we chose, and the data we considered. To make this clear in the abstract, we changed these sentences as follows.

> models, the aim of this study is to assess how equifinality affects the variability of predictions made by behavioural soil organic matter models. The results show that ~~the~~ for the model used in this study, the number of identifiable parameters, those that do not compensate for one another, increases with the number of calibration constraints~~, but~~. However, this number remained limited to five even under the most data-rich ~~conditions. Furthermore, the size~~ scenario considered, including the C and N content of particulate organic matter (POM) and mineral-associated organic matter (MAOM), and their ^{14}C signatures. Furthermore, the size of POM and MAOM can only be accurately simulated when data on these pool sizes are used, while the turnover rate of MAOM is reliably simulated only when Δ^{14}C data for MAOM are provided. Regardless of the type of mathematical equations

To make it clear that the "most data-rich" scenario applies to our study, we added this sentence to section 2.6, where the data scenarios are described.

> that in addition to total C and N, also microbial biomass C and POC were measured. Also for the SOM model, three scenarios of data availability were tested: (1) data on total SOC and N, (2) data on POC, MAOC and their C:N ratios, and (3) data on POC, MAOC, their C:N ratios, and the $\Delta^{14}C$ values of POC and MAOC. The choice of these scenarios was based on common practices in SOM modelling, although data availability may be more extensive in other studies. As the identifiability analysis

Line 35: "Three concepts that are central to achieving this balance are identifiability, equifinality and overparameterisation." The phrasing "achieving this balance" may be misleading, as identifiability, equifinality and overparameterisation are conceptual tools to describe the problem rather than mechanisms to achieve the balance. This could be clarified.

Thanks for bringing this to our attention. We fully agree with this, and with similar feedback related to this we received from the other reviewer, we changed this part to:

> Three concepts that are ~~central to achieving this balance~~ valuable to identify the mismatch between model complexity and data availability are identifiability, equifinality and overparameterisation. The concept of identifiability has a long history in scientific research (Rothenberg, 1971; Reiersøl, 1950), with different aspects of identifiability being defined and studied (Travis and Haddock, 1981; Delforge, 1977; Cobelli et al., 1979; Kleissen et al., 1990; DiStefano and Cobelli, 1980). The two aspects of identifiability that have been most frequently applied to environmental models are structural identifiability and practical

Line 135-141: If these "measurements" are artificially generated, the assumptions should be justified with references.

Thanks for pointing this out. The assumptions about the measurements were made based on meta-analyses of these properties. Now, we made it clear that our aim was not to simulate a soil in a specific location under a specific land use and provided references for studies justifying the numbers.

> In this study~~is~~, artificial SOC data were used to calibrate model parameters. We created "measurements" of the total SOC stock and fractions for one point in time, to mimic the common assumption of an SOC stock in steady state. The intention was not to replicate a soil at a specific location under a specific land use, but to use reasonable values measured across different land uses. The total SOC stock down to 0.2 m depth was calculated assuming an SOC concentration of 2 % and a bulk density of 1.2 g cm^{-3} (Chen et al., 2024; de Brogniez et al., 2015), resulting in an SOC stock of 4,800 g C m^{-2} down to 0.2 m. It was assumed that 25 % of SOC is POM and microbes in the rhizosphere (i.e., 1,200 g C m^{-2}, Hansen et al. (2024); Lugato et al. (2021)), which was divided into 96 % POC (1,152 g C m^{-2}) and 4 % microbes (48 g C m^{-2}). The remaining SOC was assumed to be ~~MAOC (~~mineral-associated organic carbon (MAOC, 3,600 g C m^{-2}). The standard deviation of the SOC pools was assumed to be 10 % of their size.

Line 136: "In this study is, artificial SOC data were .... ". "is" should be removed?

That is correct, thanks a lot for spotting this typo.

Line 140-141: MAOC should be defined at first occurrence in the main text.

Thanks for pointing this out, this has been corrected

Figure 1. Differences between rhizosphere and SOM models are unclear, and Abbreviations (MAOC, DOM, POM, MIC) are not fully defined.

Thanks for making us aware of this, we agree that the differences between both models could be better indicated. Therefore, we now show a separate conceptual figure for both models. We also explain all abbreviations in the figure caption.

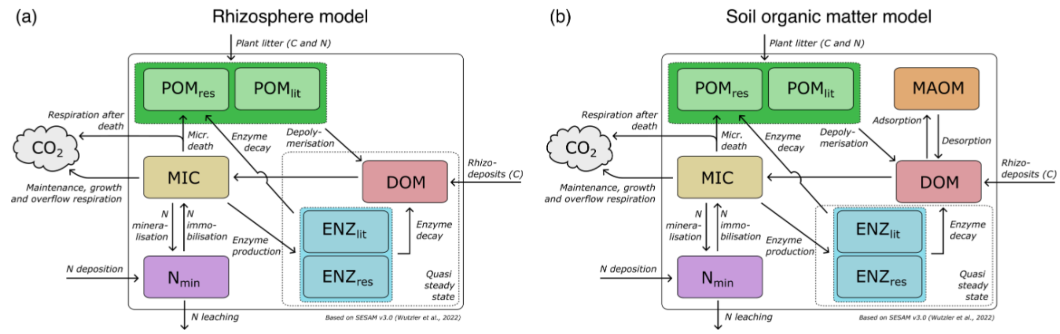

Figure 1. Illustration of the rhizosphere model (a) and the soil organic matter (SOM) model (b). The rhizosphere model is the same as the SOM model, but without the simulation of the MAOM pool. In the rhizosphere model, *DOM* is not explicitly simulated, while this is done for the SOM model. The rhizosphere model is the SESAM v3.0 model (Wutzler et al., 2022). The abbreviations are as follows: POM_{res} and POM_{lit} are particulate organic matter consisting of microbial residues and deactivated enzymes, and plant litter, respectively, *MIC* is microbial biomass C and N, N_{min} is mineral N, ENZ_{res} and ENZ_{lit} are enzymes mediating the depolymerisation of microbial residues and plant litter, respectively, *DOM* is dissolved organic matter (C and N), and *MAOM* is mineral-associated organic matter (C and N).

Figure 2: The occurrence of infinite values ("inf") is not explained. It would be helpful to clarify under which conditions these values arise and how they should be interpreted. In addition, the "inf" values are currently plotted at the boundary of the figure, which is not ideal for interpretation. A clearer representation would improve readability.The same comment applies to Figure 4.

This was indeed not explained in the manuscript, thank you very much for pointing this out. The collinearity index (γ) equals infinity if the columns of the sensitivity matrix are linearly dependent (as explained in the article describing the methods; https://doi.org/10.18637/jss.v033.i03, p. 13). This means that their values in the model perfectly compensate for each other. This has now been added to the caption of Figure 2 and 4. In addition, we have put vertical whitespace between the Inf values and the upper border of the graph.

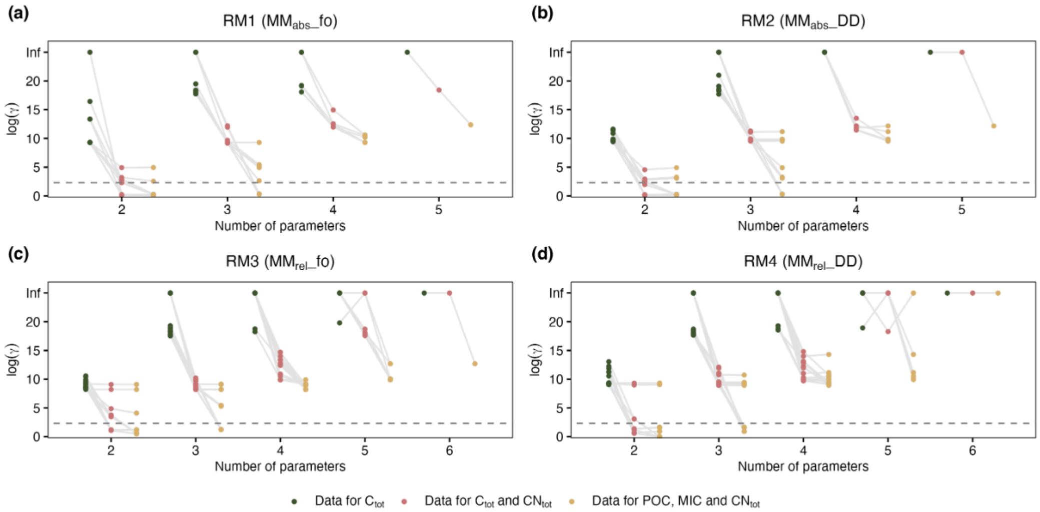

Figure 2. Logarithm of the collinearity index (γ) for all parameter combinations of the four rhizosphere models, for the scenarios where (1) only data on total SOC were available (green dots), (2) data total SOC and N were available (red dots) and (3) data on total SOC, N and microbial biomass C were available (yellow dots). The same parameter combinations are connected with grey lines. The dotted horizontal line shows the threshold in γ (log(10) = 2.3) below which parameter combinations are considered to be identifiable. Infinite values (Inf) of γ indicate that the columns of the sensitivity matrix are linearly dependent, meaning that parameter values perfectly compensate for each other. The identifiable parameter combinations are shown in Tables S11 - S18.

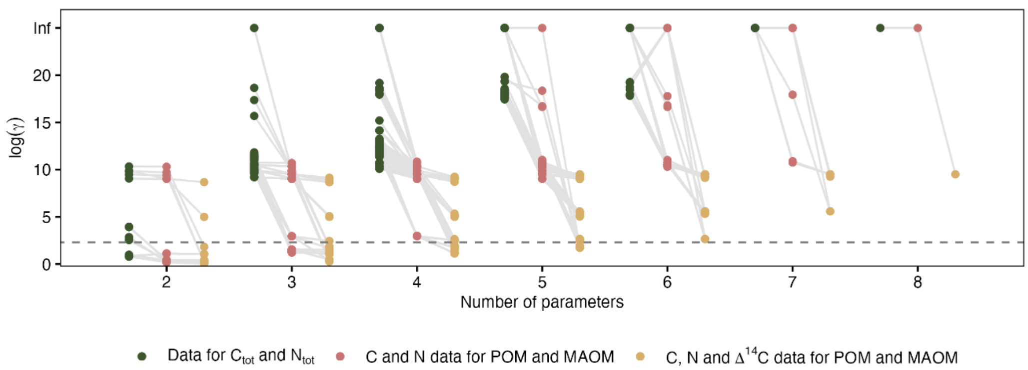

Figure 4. Logarithm of the collinearity index (γ) for all parameter combinations of the SOM model, for the scenarios where (1) only data on total SOC and N were used as calibration constraints (green dots), (2) data on the OC and N content of the POM and MAOM fractions were used (red dots), and (3) data on the OC and N content and $\Delta^{14}C$ value of the POM and MAOM fractions were used (yellow dots). The same parameter combinations are connected with grey lines. The dotted horizontal line shows the threshold in γ (log(10) = 2.3) below which parameter combinations were considered to be identifiable. Infinite values (Inf) of γ indicate that the columns of the sensitivity matrix are linearly dependent, meaning that parameter values perfectly compensate for each other. The identifiable parameter combinations are shown in Tables S19–S21.

Line 400: "This difference is due to the formulation of the rate modifiers for depolymerisation". If I understand correctly, the different non-linear equations lead to differences in decomposition rates through the "rate modifier" part. However, these differences ultimately reflect variations in the effective decompo sition rates or residence times of carbon in different pools under increased C inputs. While the current explanation *in terms of rate modifiers is mathematically correct, it would be more intuitive to present or discuss the corresponding decomposition rates or residence times of the pools.*

Thanks for this suggestion, this will indeed improve the interpretation for the readers. To do so, we have added a graph of the turnover time of the POMlit pool to figure S1.

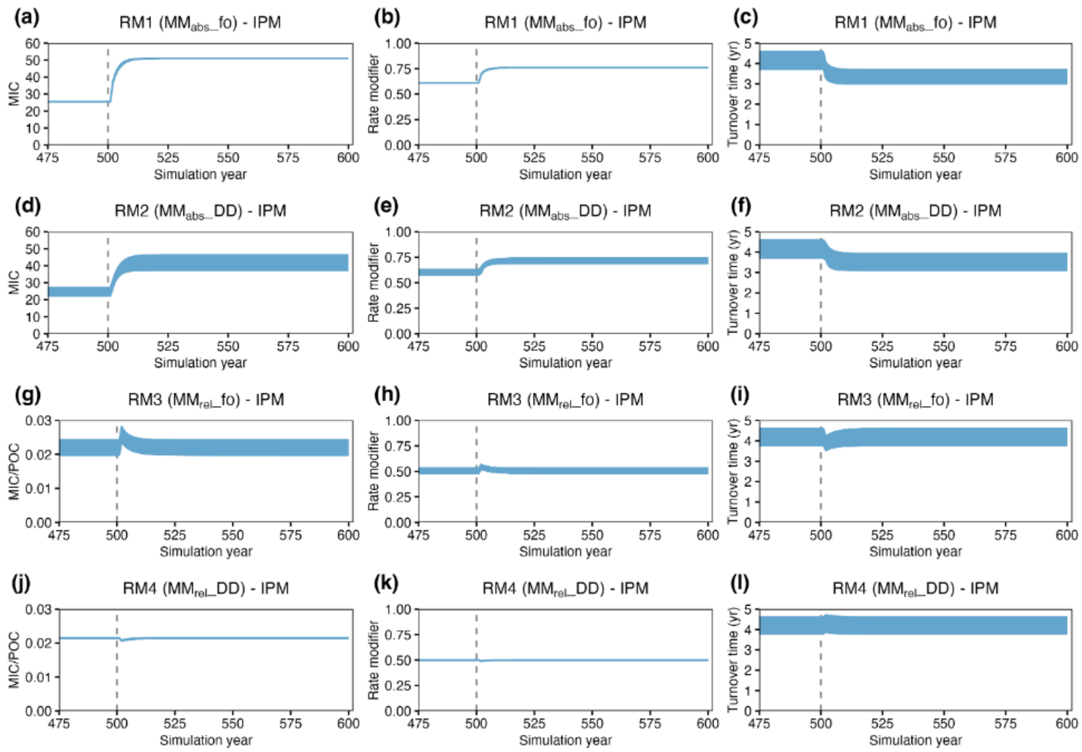

Figure S1. The effect of doubling OC inputs on changes in the rate modifiers for the depolymerisation of POC_{lit} in the rhizosphere models. Left column: model pools governing the magnitude of the rate modifiers (MIC for RM1 and RM2; MIC/POC for RM3 and RM4) before and after a doubling of OC inputs. Middle column: changes in the value of the rate modifier before and after a doubling of OC inputs. Right column: changes in the turnover time of the POC_{lit} pool before and after a doubling of OC inputs. The vertical dashed line at simulation year 500 shows from when the doubling of OC inputs took place.

We also added a brief interpretation of this to the results section.

pool, which doubled upon a doubling of OC inputs for RM1 and increased by 67.2–70.7 % for RM2 (Fig. S1 a,c). This caused an increase in the rate modifier for depolymerisation of POC_{lit} upon a doubling of OC inputs (Fig. S1 b,d), leading to a higher fraction of POC_{lit} being depolymerised compared to the initial OC inputs. This is also reflected in a decrease in the turnover time of POC after C inputs are doubled, resulting from the faster turnover (Fig. S1 c,f). In contrast, the pools controlling the value of the rate modifier for POC_{lit} in RM3 and RM4 (the ratio of MIC to POC_{lit}) remained constant upon a doubling of OC inputs, causing the rate modifier to be identical before and after a doubling of OC inputs (Fig. S1 e–h). As a result, the same fraction of POC_{lit} was depolymerised per time step upon a doubling of OC inputs~~.~~, as is also evidenced by the lack of a change in the turnover rates of POC after OC inputs were doubled (Fig. S1, i,l). This difference shows the effect of mathematical formulations on model predictions, using either absolute or relative Michaelis-Menten rate modifiers.

Figure 3: the figure lacks sufficient clarity to distinguish between MIC and full parameter model results. In particular, the large differences in prediction range (42× for RM1 and 7× for RM2) are not readily apparent.

Thanks for pointing this out. In Fig. 3, the evolution of MIC is difficult to see because the value of this pool is much lower compared to POC. We added to the caption that the evolution of the MIC pools is better shown in Fig. S1.

Figure 3. The effect of parameter identifiability and resulting equifinality on the prediction of SOC by the four rhizosphere models (RMs). Each row shows the results for a different model (see Table 1). The first column shows the results from a Bayesian optimisation of two identifiable parameters (the Identifiable Parameter Model: IPM), while the second column shows the results from a Bayesian optimisation of five (RM1 and RM2) or six (RM3 and RM4) parameters (the Full Parameter Model: FPM). Note that the lines for total SOC are not shown, as these overlapped with *POC*. The evolution of the MIC pool for the IPM models is enlarged in Figure S1. The last column shows frequency diagrams of the relative increase in SOC after a doubling of OC inputs after simulation year 500 (as shown by the vertical dashed line). The text reports the median increase, with the range between square brackets. The black dots show the average OC measurement in simulation year 500, and vertical black bars show the standard deviation.

We also clarified to the reader how these differences in prediction range (42 and 5 times) can be seen in the figure.

Also the number of optimised parameters had an effect on predictions by the rhizosphere models. For RM1 and RM2 (with absolute Michaelis-Menten kinetics), the average relative increase in OC upon a doubling of inputs was lower for the full parameter models than for the identifiable parameter models (Fig. 3). In contrast, for RM3 and RM4 (with relative Michaelis-Mention kinetics), overparameterisation did not affect the average increase in OC, which was ca. 100 %. Moreover, the range in predictions was ca. 42 times larger for RM1 for the full parameter models (a range of 34 %, i.e., between 34.5 and 68.5 %, Fig. 3c) compared to the identifiable parameter models ~~, the~~ (a range of 0.8 %, i.e., between 61.1 and 61.9 %, Fig. 3c). For RM2, the range in predictions was ca. 6 times larger for the full parameter models ~~showed a prediction range approximately 42 times larger for RM1 and 7 times larger for RM2.~~ (a range of 21.2 %, i.e., between 54 and 75.2 %, Fig. 3f) compared to the identifiable parameter models (a range of 3.5 %, i.e., between 67.2 and 70.7 %, Fig. 3f). These results show that the accurate prediction of steady-state stocks of SOC by behavioural models is not a sufficient criterion to evaluate the predictive capabilities of such models.

Section 3.2.1: It is unclear which figures support the results described in this section, as no figure is explicitly cited. Given that this is part of the Results section, appropriate figure references should be provided to support the statements. In addition, Figure 4 is not clearly introduced or discussed in the Results section before the manuscript proceeds to Figure 5. This disrupts the logical flow of the results. Please ensure that all figures are properly introduced and described in sequence.

Thanks for pointing out this oversight. This section is based on the results presented in Fig. 4, which is now cited in this section to improve the flow of the manuscript.

> The parameter identifiability analysis for the SOM model showed that, similar to the results for the rhizosphere models, the number of identifiable parameters increased with an increasing quantity of calibration data ~~.~~(Fig. 4). When only data on total SOC and N were used, at most two parameters were jointly identifiable, while three parameters were identifiable together when

Figure 5: The color scheme is difficult to interpret, particularly because MAOC and MIC appear very similar, and the meaning of some colors (e.g., dark orange) is unclear. This also makes it difficult to support the statement in Lines 427–429 that SOC can be dominated by either POC or MAOC (Fig. 5b), as the contribution of POC is not clearly visible. Improving the color scheme and legend would enhance clarity and consistency between text and figure.

Thanks for noticing the similar colors between MAOC and MIC in the legend. This was caused by plotting the lines transparent, which also caused the legend colors to be transparent. This has been fixed now, so that it's clear from the legend which colors are for MAOC and MIC. We note that also in this plot, the lines for MIC are very difficult to see because of the small size of this pool. We clarified this in the caption.

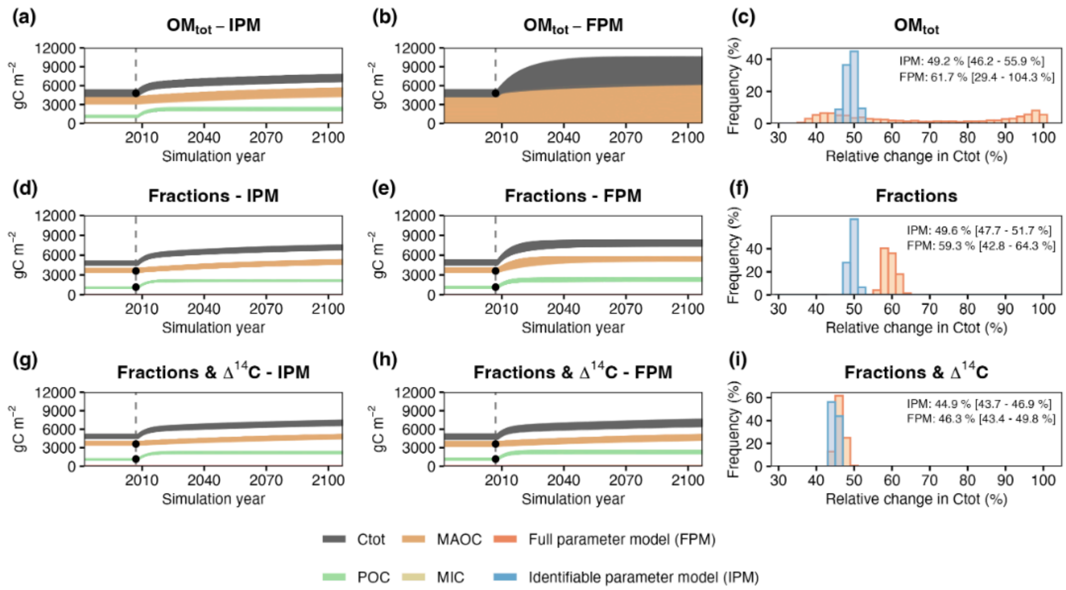

Figure 5. Results of the Bayesian calibration of the SOM model using different data constraints: (1) only total SOC and N (*OM_{tot}*; a–c), (2) the OC and N content of the POM and MAOM fractions (*Fractions*; d–f), and (3) the OC and N content and Δ^{14}C value of the POC and MAOC fractions (*Fractions & $\Delta^{14}C$*; g–i). The first two columns show the simulations for the different simulated pools, with the first column showing the results for the case when only identifiable parameter sets were optimised (the identifiable parameter model: IPM), while the second column shows results for the case when a non-identifiable parameter set (i.e., all model parameters) was optimised (the full parameter model: FPM). We note that due to the small size of the *MIC* pool, these lines are difficult to see. The black circles show the data used for parameter optimisation, while the vertical dashed lines show the timing of the doubling of OC inputs. The corresponding simulations of Δ^{14}C are shown in Fig. 6. The histograms on the right show the relative change in SOC for the year 2107, after a doubling of OC inputs from the year 2007 onwards. The text reports the median increase, with the range between square brackets.

Regarding the statement that SOC can be dominated by either POC or MAOC (Fig. 5b), we now made it clear that the lines for POC are covered by the lines for MAOC. This is unfortunately inevitable, as both POC and MAOC compensate for each other, and span the entire range of potential values from a very small pool size to almost all SOC present in these pools.

calibration scenario had a substantial impact on the internal dynamics of the model. For example, when model parameters were optimised using only data on total SOC and N while calibrating all model parameters, the majority of SOC could be either in *POC* or *MAOC* (Fig. 5b, note that the lines for the *POC* pool are covered by the lines for the *MAOC* pool, as both pools span the entire range of potential values from a very small size to almost all SOC present in these pools; their density distribution is shown in Fig. S3), clearly demonstrating equifinality resulting from overparameterisation. Further evidence for

To make this better clear to the reader, now we also show a plot of the density distribution of the amount of POC and MAOC at steady state for this model in the supplement, and refer to this plot in the main text (see textbox above).

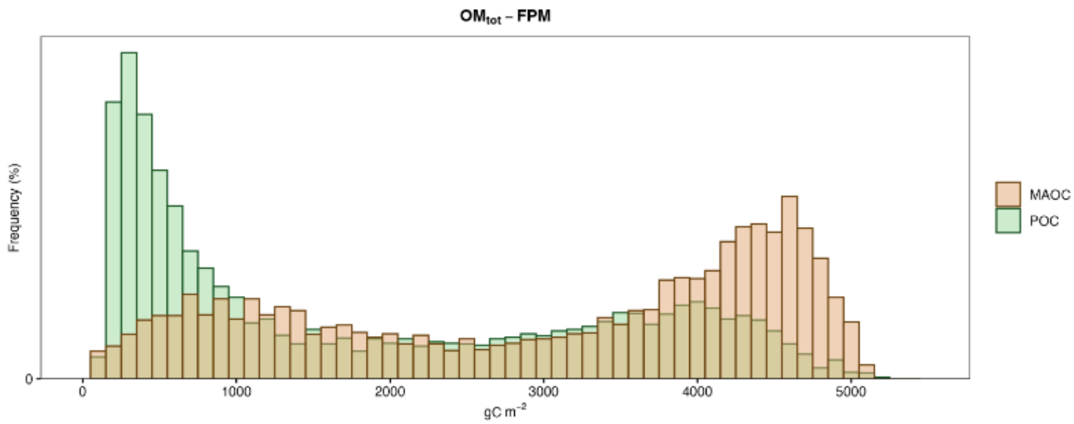

Figure S3. The simulated size of particulate organic carbon (POC) and mineral-associated organic carbon (MAOC) pools for the behavioural SOM models at steady state (i.e., the last year before OC inputs were doubled) for the SOM model optimised using only data on total SOC and N while calibrating all model parameters (full parameter model; FPM).