Reviewer 1

1. In their manuscript, titled 'Towards resolving poor performance of mechanistic soil organic carbon models', Wang et al. describe their results of a study comparing the performance of two mechanistic soil organic carbon models (MIMICS and the newly proposed MES-C) to a machine learning approach (random forest). They use the mechanistic models and random forest models to predict the amount of soil organic carbon for multiple locations around the world, after which they compare model results to depth profiles of SOC from the WoSIS database. Based on their results, the authors conclude that random forest models consistently outperformed mechanistic models. In addition, they used interpretable machine learning methods to conclude that both mechanistic soil organic carbon models perform poorly due to lack of accounting for key variables (most notably CEC), while the role of existing variables is underrepresented. They compare the controlling factors in the mechanistic models to factors controlling the SOC content at the global scale, and find that, for example, the role of the clay and silt content of the soil is overestimated by the models, compared to observations. Furthermore, the authors asses the sensitivity of model outcomes along a range of predictor values, and study how pairs of predictors affect model outcomes, compared to observations.

   The development and application of mechanistic soil organic matter models, among which MIMICS, has gained much importance over the past two decades. Therefore, the assessment of model performance and the relative importance of model drivers in making predictions of the SOC content, compared to observations, is needed to improve these models to increase confidence in their outcomes. I appreciate the effort done to collect the data, develop a new model and apply it, together with MIMICS, at the global scale, which is a challenging endeavour. The manuscript is well-written and well-structured. The abstract conveys the main messages and the introduction provides sufficient background to the topic, although I would encourage the authors to include more information about previous studies assessing controls of the SOC content at the global scale, which is the topic of their study. Although the methods section describes the performed analyses and data collection strategy, much information about the performed model simulations is missing to assess the quality of the performed simulations (see below). The results section concisely describes the results, but I would encourage the authors to show the model results in a graphical way, for example using scatterplots. The discussion touches upon the main results, but mainly sections 4.3 and 4.4 lack clear messages, and are rather a summing up of the relations the authors found.

We would like to thank the reviewer for time and feedback on this manuscript. We'll revise the manuscript through according the reviewer's comments on all sections, particularly providing more information about previous studies on the dominant controls of the SOC content at global scale, a more detailed description of model structure and parameter optimisation processes, and revising the discussion to avoid restating the results but place greater emphasis on consistent findings across methods, as well as the limitations and uncertainties of this study. Please find our point-by-point responses below.

2. Based on my evaluation I recommend major revisions, mainly due to (1) the lack of detailed information which would be necessary to evaluate the performed model simulations and (2) the nearly complete attribution of the mismatch between model results and measurements to the structure of the model, without accounting for

uncertainties in input data at the global scale (see below). My main concerns about the manuscript are the following, while detailed feedback is provided below.

2.1 The authors propose a new mechanistic soil organic matter model that is more complex than MIMICS, as it also incorporates SOC in aggregates. However, the authors do not have any data on the distribution of measured SOC among different pools, so they have no data to either constrain the distribution of simulated SOC among the different model pools, or to evaluate this simulated distribution. Therefore, I invite the authors to justify the use of such a (complex) model at the global scale, given this lack of data. The evaluation of the simulation results based on data for only total SOC inevitably leads to overparameterization and equifinality. This is an important, but generally overlooked, aspect of environmental models, the consequences of which for the model simulations should be discussed in the manuscript.

Thanks for the insightful suggestion. We acknowledge that the lack of observed carbon fraction data prevents us from directly constraining and validating the simulated dynamics of SOC pools, which inevitably introduces the risk of overparameterization or overfitting in this study. Nevertheless, we consider it important to incorporate more mechanistic processes such as aggregation and adsorption/desorption because their roles in long-term stabilization of SOC are well studied and documented. Our new model provides the framework that can be directly tested once the global SOC fraction data are collected, a direction we're actively pursuing now. We'll revise the Discussion to highlight these limitations and uncertainties regarding the SOC pool distributions.

2.2 The authors attribute the 'poor performance' of the mechanistic models exclusively to the structure of these models. However, there are many more uncertainties leading to a mismatch between model results and measurements:

- Uncertainties related to model drivers, which were extracted from global data sets. For example, NPP data derived from MODIS cannot be expected to correctly represent C inputs from vegetation to the soil, as, for example, a substantial part of aboveground litter C will be respired before it can enter the soil. A model can never be better than the input data, so accounting for uncertainties in the data is important.

- The applied mechanistic models are too complex compared to the data available for model evaluation, i.e., they are overparameterized. Six parameters are evaluated while only one output (total SOC) is evaluated (it seems, as this is not described clearly in the manuscript). This inevitably leads to large uncertainties about model results.

These aspects should be discussed, because it is evident that an overparameterized model with uncertain inputs will lead to a 'poor performance'.

Thanks for pointing out additional sources of model uncertainty. In this study, we focused mainly on the weakness of model structures relating to how well the models capture the relationships between SOC and environmental drivers globally. We randomly selected 80% of observations (training data) to optimise parameters in process-based models and tune two random forest models, while the remaining 20% of observations (test data) were used for validation. This procedure was repeated 10 times for cross-validation. The uncertainty in

drivers such as NPP is indeed considerable, and using machine learning alongside process-based models is one way to assess how problematic this issue is. Additional details about this will be provided in the revised manuscript, including further discussion on uncertainties in model drivers and issue of overparameterization.

The methods related to the application of the mechanistic models are not sufficiently described. For example:

- Down to which depth were the simulations performed? To 1.5 m, as this is the maximum depth down to which measurements were available?

Yes, the simulations in this study were conducted down to 1.5 m. While WOSIS includes SOC observations below this depth, these data are highly uncertain, and the explained variance in deeper layers is very low, even when using Random Forest with 12 predictors. For this reason, we did not consider it necessary to include deeper layers in the present analysis. Importantly, our study provides a general framework for applying multiple explainable artificial intelligence approaches to diagnose process-based models, which can be extended to deeper soil layers once reliable SOC observations become available. We'll of course make this information clearer in the revised manuscript.

- Were the simulations performed at a vertical resolution of 30 cm?

The vertical resolution in process-based models is 10 cm, but we aggregated the outputs to 30 cm using the mean value of three layers to compare with observations. We'll describe this clearly in the revised manuscript.

- For how long were the simulations run? Was there a spin-up period, followed by a run with actual temperature and precipitation data?

In this study, we assumed that SOC are at equilibrium, and we ran the models for around 1000 years until equilibrium was reached. This corresponds to a spin-up simulation, but we did not include a subsequent transient run with annually varying climate inputs. We'll revise our manuscript to include this information.

- In which programming language were the models coded?

Both MIMICS and MES-C were written in Fortran and optimised for GPU. We'll add this information to Section 2.1 Model description.

- With which time step were the simulations performed? Which solver was used to solve the differential equations?

Both MIMICS and MES-C were run with an hourly time step, and the differential equations were solved using the fourth-order Runge–Kutta (RK4) method. We'll provide more detailed descriptions of both process-based models in the revised manuscript.

- How were locations with a different land use treated? Was there a different vertical resolution for C inputs per land use?

We did not explicitly consider land-use in this study. SOC for all sites was simulated at a vertical resolution of 10 cm, with the vertical allocation of carbon inputs determined by a negative exponential function describing the decline of root distribution with depth (Jackson et al., 1996),

$$Y = 1 - \beta^d$$

Where Y is the cumulative root fraction from the soil surface to depth d (cm), and β is the fitted parameter. In this study we applied the same β (0.966) for all vegetation types. We acknowledge that this approach may introduce uncertainties in SOC predictions at depth. However, due to the lack of consistent parameters describing root distributions for different vegetation types, and because approximately 70% of roots are generally concentrated in the top 30 cm of soil across most vegetation types, we expect the resulting uncertainties to be limited, particularly in the top 30 cm, which is the main focus of our study. We'll revise our manuscript to include this information.

2.3 No sufficient information about model evaluation is provided. As the authors conclude that the mechanistic models perform poorly compared to machine learning approaches, a thorough model evaluation is important to support this conclusion. For example:

- What was the simulated turnover time of SOC at different soil depths? Was this in line with measurements (e.g., DOI: 10.1126/science.aad4273) Did this decrease with soil depth, as is generally observed?

In our simulations, the mean turnover time of total SOC increased with depth, ranging from ~60 years in surface soils to ~800 years in the deepest layer, while the turnover time of mineral-associated organic carbon (MAOC) exceeded 2000 years at deep depths. This pattern of increasing turnover time with depth is consistent with general observations, although the absolute values we report are shorter than radiocarbon-derived estimates (e.g., He et al., 2016). This discrepancy arises because our model was not constrained by radiocarbon data, which is known to capture the persistence of older carbon pools more accurately. We will clarify this limitation in the Discussion and highlight the associated uncertainties.

- How was the simulated amount of SOC distributed among the different pools? Was this distribution in line with observations?

The mean simulated proportion of MAOC in total SOC across sites is ~40% in the top 30 cm, which is lower than the mean proportion reported in the LUCAS dataset (~60%; Cotrufo et al., 2019), but still falls within the observed range of that dataset. Importantly, the model also reproduces the expected increase in MAOC proportion with depth, consistent with established patterns of vertical SOC stabilisation. We acknowledge that our model behaviour is not yet fully constrained due to the limited availability of global observations of SOC fractions. Nevertheless, our framework provides a testable model structure that can be further evaluated and improved as more comprehensive SOC fraction datasets become available, a direction we are actively pursuing. We'll discuss more about the limitation and uncertainties of the model in the revised manuscript.

3. The authors should justify why they chose to perform a global study, instead of focusing on a smaller region with less uncertainties about input data for the models and more detailed data for model evaluation.

Applications of process-based SOC models at regional and continental scales have generally achieved good performance (Zhang et al., 2020; Wang et al., 2024), whereas at the global scale their performance remains poor. Our motivation for conducting a global study is to address this gap. In addition, SOC dynamics are a critical component of Earth System Models (ESMs), which operate at the global scale. Improving the performance of SOC models globally is therefore essential to enhance their incorporation into ESMs and to improve predictions of carbon–climate feedback. We'll add this information to the Introduction.

Detailed feedback

Abstract

L23: Would be good to mention which 'existing variables' these are

Thanks for the suggestion, we'll describe the existing variables explicitly in the Abstract.

L29: What do you mean with a 'simple trend'?

It means 'monotonic' here, we'll revise this to be clearer.

Introduction

L45-55: It is not clear what the 'effectiveness of land management strategies' has to do with the ability to accurately estimate SOC content

Sorry for the lack of clarity here. Our intention was to highlight that implementing effective land management strategies for SOC sequestration requires reliable baseline estimates of SOC stocks and their spatial distribution. Without accurate knowledge what controls the current SOC levels, it is difficult to quantify additional SOC storage attributable to management interventions, or to assess their potential contribution to offsetting CO_2 emissions. We will revise the text to clarify this link explicitly.

L73: The study by Gurung et al. is a study on daycent, so not applicable as a reference for the 'poorly constrained parameter values of microbial explicit models'

Thanks for pointing this out. We'll delete the reference here.

L83: Georgiou et al. (2022) is not a modelling study, so not appropriate as a reference for this sentence.

Thanks for pointing this out.

We'll replace it with Abramoff et al., 2022 (https://doi.org/10.1016/j.soilbio.2021.108466) and Laub et al., 2024 (https://doi.org/10.5194/gmd-17-931-2024).

L107: 'accuracy' of what?

Sorry for the confusion here, we'll change it to 'SOC prediction accuracy'.

L111: R^2 is not a measure of model performance (see e.g. https://en.wikipedia.org/wiki/Anscombe%27s_quartet), so please remove this throughout the manuscript when used for this purpose.

Thank you for this suggestion. We agree that R^2 alone is not an adequate measure of model performance, as it reflects explained variance rather than prediction accuracy. In our study, however, one of the main objectives is to assess how much of the variation in global SOC can be explained by different models and to explore ways of improving the variance explained by process-based models. For this reason, we have retained R^2 as a measure of explained variance. At the same time, we have used additional metrics including RMSE, MAE, and AIC (see Table 2) for evaluating model predictive performance. We also recognize the importance of directly assessing model–data agreement, and in response to this comment, we will include Lin's Concordance Correlation Coefficient (CCC) in the revised manuscript to provide a more robust measure of the agreement between simulated and observed SOC values. Including both R^2 and other complementary statistics allows us to evaluate models comprehensively while also facilitating comparison with previous studies in this field, where R^2 is widely reported (e.g., Chen et al., 2024; Nyaupane et al., 2024).

Methods

General methods: Does the WoSIS database contain SOC stocks, i.e., a combination of OC% and bulk density, for all profiles that were used? If not, how was OC% converted to SOC stocks? Although it is not described in the manuscript, I assume absolute amounts of SOC were simulated? To compare these to observations of OC% in WoSIS, these would need to be converted to OC%. How did this happen? This information should be provided in the manuscript.

SOC in WOSIS is reported in g C kg^{-1} soil. Both MIMICS and MES-C simulate SOC in mg C cm^{-3} soil, and these values were converted to g C kg^{-1} soil using bulk density data from SoilGrids2.0. We will add this clarification to the Data and Methods section.

L143-144: where did the values for the parameter to distribute litter inputs over metabolic and structural litter come from? As this is based on litter quality, was this value different for different types of vegetation?

The partitioning of litter to metabolic litter is dependent on litter lignin and nitrogen (N) content (Wieder et al., 2015),

$$0.85 - 0.013 \times (lignin/N)$$

In our study, we parameterized this function using assumed litter quality values. Specifically, we set lignin/C ratios to 0.15 for grasses and 0.25 for woody plants, respectively. We set C/N ratios to 35 for aboveground litter and 60 for belowground litter, respectively. And $lignin/N$ ratios were then calculated as the product of lignin/C and C/N. These values differ between vegetation types (grasses vs. woody) and litter origin (above- vs. belowground) to reflect in litter quality. We'll add this information to section 2.1 Model description.

L143-158: more information about the chosen version of MIMICS is needed. 1) What did the soil moisture scalar look like? This is important information, as you discuss the limited ability of MIMICS to account for moisture later in the manuscript. 2) How was bioturbation simulated? 3) Was the magnitude of diffusion the same for every land use and vegetation type?

4) Were the same parameter values used for all simulated depths? 5) How was, for example, the generally-observed decrease in the rate of SOC cycling with depth accounted for?

Thanks for all questions. We'll describe both process-based models in detail in the revised manuscript.

1) Soil moisture scalar in this study follows Ghezzehei et al., 2019 and is described in the Supplementary Material.

2) The vertical transfer of SOC via bioturbation and diffusion we used in this study was described in Wang et al., 2021 (https://doi. org/10.1029/2020JG006205).

3) We didn't consider vegetation types explicitly in this study. However, the diffusion magnitude is different for environmental clusters (see Section 2.3) by tuning the diffusion coefficient (see Table S2).

4) Yes, the same parameter values are used for all depths, except that both microbial maximum reaction rate (V_{max}, see below) and litter carbon input decrease with soil depth.

5) The depth-dependent decline in SOC turnover was following Koven et al. (2013), using an exponential decay function,

$$r_z = \exp\left(-\frac{z}{z_\tau}\right)$$

Where r_z is depth-dependent scalar, z is soil depth (cm) and z_τ (0.5) is the e-folding depth of intrinsic turnover rates. This scalar was applied to the microbial maximum reaction velocity (V_{max}) to reduce SOC turnover rates with increasing depth.

L181-182: In MIMICS, doesn't the physically-protected SOC pool represent mineral-associated OC?

In MIMICS, the physically-protected SOC pool includes both aggregate-protected carbon and mineral-associated carbon, and its turnover reflects contributions from aggregate breakdown as well as mineral desorption.

L183: this knowledge is not 'recent' but has been known for decades

Thanks for pointing this out. We'll change this text to "to integrate more process-based SOC stabilisation theories".

Section 1.2.1: as you describe a new model (although based on MIMICS), I encourage the authors to describe the equations that they added to MIMICS in the main manuscript, not only in the supplement.

Thank you for this suggestion. The primary focus of our manuscript is on applying multiple artificial intelligence methods to identify and analyse potential weaknesses of process-based models in representing environmental controls on SOC, rather than on model development itself. Since MIMICS is already a published model and the modifications we introduced have been documented in detail elsewhere, we prefer to keep the full set of equations in the Supplementary Materials. However, we will expand the description in the Supplement Material to ensure clarity and provide sufficient details for reproducibility.

Section 2.2.1: did you check for autocorrelation between the predictors before performing the analyses?

Yes. Prior to the analyses, we examined pairwise correlations among predictors and retained only those variables with correlation coefficients below 0.7, thereby reducing potential issues of collinearity.

L223-225: did this 'negative exponential function' have different parameters for different vegetation types?

We applied the same parameter for all sites. Due to the lack of consistent parameters describing root distributions for different vegetation types, and because approximately 70% of roots are generally concentrated in the top 30 cm of soil across most vegetation types, we expect the resulting uncertainties to be limited—particularly in the top 30 cm, which is the main focus of our study. We'll revise our manuscript to include this information.

L233: weighted based on what?

Soil properties are weighted based on soil thickness. We'll revise the sentence to make it clear.

L234: what do you mean by 'standardized'?

We will replace "standardized" with "harmonized" for clarity. Soil temperature and moisture from ERA5-Land are originally reported for the 0–7, 7–28, 28–100, and 100–289 cm layers. To obtain values at uniform 30 cm intervals, we calculated a weighted average for each 30 cm layer based on the thickness of the original layers.

L235: 'mean annual values': over which time period where these means calculated?

The time periods covered by the different variables vary according to data availability, and details are provided in Table 1.

L248: what is 'SOC profile data'? Only OC%, or also bulk density, to convert OC% into stocks?

Here we refer to only OC% in WOSIS.

L255: so all profiles that were used had data down to 1.5 m? Please clarify this in the manuscript

Sorry for the confusion here. We'll revise it to "SOC values were harmonized to 30 cm intervals down to a maximum depth of 1.5 m using a spline function". Not all profiles have observations down to 1.5 m, please see L266 – L267 and Figure 2b for details.

L257: what happened to 1x1 km grid cells that did not have any profiles in WoSIS? Were these cells not simulated?

Yes, they were not simulated.

L262: It is not clear how a cumulative SOC content (in g/kg) can be calculated, as you can only calculate a cumulative profile by summing absolute values, not concentrations. Please clarify.

We apologize for the incorrect unit for SOC here, it should be carbon density (kg m^{-2}). SOC in WoSIS is reported as concentration (g C kg^{-1} soil), which we converted to carbon density using soil layer thickness (30 cm in this study) and bulk density. Although bulk density is recorded in WoSIS, it is missing for many profiles, so we used SoilGrids as a replacement. Since SoilGrids bulk densities were generated using WoSIS observations as training data, we consider them reliable for this purpose.

L266-267: 1) If no data below 60 cm was present, where these deeper layers also simulated? 2) How did you resample the data to 0-30, 30-60, … intervals? 3) Were OC% weighted based on bulk density? Or were averages of the OC% used? Please clarify.

1) Deeper soil layers were simulated but excluded from the calculation of cost function during parameter optimisation.

2) SOC data were harmonized to 30 cm intervals using a cubic spline interpolation from *ithir* package (Malone et al., 2017) in R.

3) Process-based models simulated SOC in mg C m^{-3} for each 10 cm depth up to 1.5 m. We calculated SOC concentration (g C kg^{-1} soil) using bulk density for corresponding depth, and then the average of SOC concentration of each 30 cm soil layer is used to compare with observations.

Fig. 1: Why is there a sharp boundary latitude below which no data are present?

The original WoSIS SOC profile dataset has global coverage. However, after applying our selection criteria (see Section 2.2.2), only a subset of profiles was retained. Many profiles in South Hemisphere were excluded due to insufficient depth observations, or because some environmental variables or ancillary datasets required for model parameterisation were missing.

Section 2.3:

- More information about the parameter optimization process is needed:
    - Which data were used to optimize model parameters? Only total SOC%? If so, how can you assess if the distribution of simulated SOC over the different model pools was correct?

Yes, we optimized the model using only total SOC. Due to the lack of globally consistent SOC fraction datasets, we did not impose explicit constraints on the distribution of SOC among different pools during parameter optimisation. However, we evaluated the simulated pool distributions post hoc and adjusted parameters related to carbon allocation to ensure that the fractions of different pools (e.g., MAOC, microbial carbon) fell within ranges reported in published studies. This approach allowed us to maintain consistency with empirical knowledge, even though direct global constraints on SOC fractions were not available in this study.

Which measure was optimized? RMSE?

RMSE is optimized in the cost function, please see L288-L289 for details.

    - Which program was used to optimize model parameters?

We used a global optimisation algorithm called the shuffled complex evolution (SCE-UA, version 2.2) method (Duan et al., 1993) written in Fortran. Please see L286–L289 for details.

- Did you use one constant desorption parameter for the entire globe, or per 'cluster'? If so, the turnover rate of SOC is very unlikely to be simulated correctly, as this is different in different ecosystems/soil types/… Please clarify.

In this study, we used a constant desorption parameter within each cluster. The clusters were defined based on similarities in both climate conditions and soil properties, which we consider providing a more representative grouping than using ecosystem or soil types alone.

Please provide the optimized values of the calibrated parameters.

Thanks for the suggestion. We'll add the values of optimised parameters in the revised manuscript.

L285: what do you mean by 'relatively sensitive'? How was this sensitivity calculated?

Sorry for the missing information here. We did a two-step parameter sensitivity analysis (Lu et al., 2013) before optimization: a screening step that ranks all model parameters by their importance on model output in order to select the potentially important parameters, and a second step that aims to quantify the contribution to the variance of model output by each of the pre-selected parameters and by their interactions. We'll add the details of parameter sensitivity analysis in the revised manuscript.

L300: what do you mean by 'for reference'

Apologies for the confusion. In the calibration, we excluded sites with mean annual temperature below 0 °C, as process-based models may not adequately capture SOC dynamics in frozen soils. However, we initially applied the calibrated model to these sites. For clarity and consistency, we will remove all sites from frozen areas in the revised manuscript.

L302-303: R2 is not a measure for model performance, and should not be used as such (see above). Please remove and perform the analysis without this measure.

Thank you for this suggestion. As we explained above, beyond comparing the performance of machine learning and process-based models, our main focus is on assessing how much of the variation in global SOC can be explained by different models and on identifying ways to improve the explained variance of SOC in process-based models. For this purpose, we prefer to retain R^2 as an indicator of explained variance. At the same time, we have used additional metrics including RMSE, MAE, and AIC (see Table 2) for evaluating model predictive performance. We also recognize the importance of directly assessing model–data agreement, and in response to this comment, we will include Lin's Concordance Correlation Coefficient (CCC) in the revised manuscript to provide a more robust measure of the agreement between simulated and observed SOC values. Including both R^2 and other complementary statistics allows us to evaluate models comprehensively while also facilitating comparison with previous studies in this field, where R^2 is widely reported (e.g., Chen et al., 2024; Nyaupane et al., 2024).

Results

Section 3.1: scatterplots (measured vs modelled) would be necessary to evaluate model performance. I encourage the authors to provide these.

Thank you for this helpful suggestion. We agree that scatterplots provide an intuitive way to assess agreement between observed and simulated SOC. However, given that we evaluate four models across five soil layers using 10-fold cross-validation, including scatterplots for all cases would result in an overly complex presentation. To address this, we will include Lin's Concordance Correlation Coefficient (CCC) in the revised manuscript, which directly quantifies the agreement with the 1:1 line. Models with CCC values close to 1 indicate strong concordance with observations, complementing the other performance metrics already reported.

L352: Clarify what you mean by 'out-of-sample'

By "out-of-sample", we mean evaluating the model on test data that were not used for training.

L357-358: how do you see in Table 2 that that the predictability decreases with depth, since all measures decrease with depth? Better would be to use relative measures (scaled by the measured OC%).

Thanks for the good suggestion. To account for the smaller magnitude of SOC in deeper layers, we will evaluate model performance using log-transformed SOC values in the revised manuscript. This effectively scales the errors relative to the measured SOC at each depth and provides a more balanced assessment of predictability across the soil profile.

Table 2: I encourage the authors to remove R2 from the table, as this is not a measure of model performance (a very bad simulation, for example a systematic over- or underestimation, can have an R2 value close to 1).

Thank you for this suggestion. As we explained above, beyond comparing the performance of machine learning and process-based models, our main focus is on assessing how much of the variation in SOC can be explained by different models and on identifying ways to improve the explained variance of SOC in process-based models. For this purpose, we prefer to retain R^2 as an indicator of explained variance, while clarifying in the revised manuscript that we do not use it as a comprehensive measure of model performance.

Fig. 3: how is it possible that in a) NPP has a ranking far below AMT, while in b) (same data but fewer predictor variables) this order is reversed?

The difference in variable importance ranking between a) and b) is due to the set of predictors included in the model. In random forest, the importance of a variable is relative and depends on how much unique variance it explains compared with other predictors. In panel a), when all 12 predictors are included, NPP is moderately correlated with some climate variables such as AP ($r = 0.6$). Because of this correlation, the random forest tends to assign the explanatory role to AP and similar variables when constructing the initial tree splits, and NPP's incremental contribution to explaining SOC variance is relatively smaller, resulting in a lower importance ranking. In panel b), with a subset of only 5 predictors, many of the variables that explained similar amount of SOC variance as NPP are removed. NPP now captures variance that was

partially explained by other variables in a), so its relative importance increases, surpassing soil temperature.

Fig. 3: Is the variable Soil Temp different from AMT? Why is this variable not present in plot a)? Please clarify.

Soil temperature and AMT are derived from different datasets, but they are highly correlated ($r > 0.9$). We used soil temperature rather than AMT as an input to the process-based model because it better represents the actual conditions experienced by soil microbes and organic matter, therefore Soil Temp is in panel b) to show the importance of model inputs on SOC. For the random forest analysis, however, we used AMT to maintain consistency with the other climate variables, as they come from the same dataset.

L547-549: While this is partially true, there is still a substantial correlation between clay+silt content and maximum potential mineral-associated SOC (see https://doi.org/10.5194/soil-10-275-2024). I encourage the authors to put some nuance to this statement.

Thank you for the suggestion. We agree that clay and silt content shows a substantial correlation with maximum potential mineral-associated SOC and is widely used in process-based models as a proxy for soil physical properties due to ease of measurement. However, clay content alone may not fully capture important variation in mineral composition, specific surface area, or surface chemistry, which critically influence SOC stabilisation and sequestration (Bailey et al., 2018). For example, soils with similar total clay content can differ markedly in SOC due to differences in clay mineral types and their sorptive capacities (Parfitt et al., 1997). We will revise the discussion to reflect these nuances, highlighting both the utility and limitations of clay and silt content as a proxy for SOC stabilization.

Discussion

L503 + 511 + 517-518: Please remove these statements or use a performance measure different than R2 (see above)

Thank you for this suggestion. As we explained above, beyond comparing the performance of machine learning and process-based models, our main focus is on assessing how much of the variation in SOC can be explained by different models and on identifying ways to improve the explained variance of SOC in process-based models. For this purpose, we prefer to retain R^2 as an indicator of explained variance, while clarifying in the revised manuscript that we do not use it as a comprehensive measure of model performance.

L534: how should this be done? A substantial part of CEC comes from organic matter itself (see for example Solly et al., which you cite), which is not discussed in the manuscript. Soils with a high SOM content typically have a higher CEC value, so it seems that including CEC as a model driver will not improve the mechanistic basis for these models. Please discuss.

Thank you for this valuable comment. We agree that a substantial fraction of CEC is derived from organic matter itself, particularly in topsoil, as noted by Solly et al, 2020. Our intention was not to suggest CEC as an independent driver in isolation, but rather to highlight its role as an integrative property reflecting both organic matter and mineral contributions to soil sorption capacity. Including CEC in models may therefore help constrain mechanisms of SOC stabilization, particularly in soils where mineral phases (e.g., clays, Fe/Al oxyhydroxides) dominate CEC contributions. We will revise the discussion to clarify this point and

acknowledge that the utility of CEC as a model driver likely depends on distinguishing between its organic and mineral sources.

Technical feedback

L50: ‘is’ => ‘was’

Thanks, we’ll amend it.

L77: ‘progress’ => ‘process’?

Thanks, we’ll amend it.

L78: ‘limited’ => ‘few’?

Thanks, we’ll replace ‘limited’ with ‘few’ here.

L88-89: something seems wrong with this sentence

Sorry for the lack of clarify here. We’ll revise it to “higher temperature stimulates soil microbial activities, which in turn accelerates SOC decomposition”.

L138: ‘variable’ => ‘variables’

Thanks, we’ll amend it.

You use AMT as an abbreviation for mean annual temperature, why not stick to the conventional MAT abbreviation?

Thank you for pointing this out. In WorldClim2, the variable is defined as Annual Mean Temperature (AMT), which is why we initially adopted this abbreviation. However, we agree that MAT is the more conventional term in the literature. To avoid confusion, we will revise the manuscript to use MAT consistently.

L299: ‘permafrost’ => ‘frozen’?

Thanks, we’ll change it to “frozen”.

L554: ‘perform’ => ‘performs’

Thanks, we’ll amend it.

References

Abramoff, R. Z., Guenet, B., Zhang, H., Georgiou, K., Xu, X., Rossel, R. A. V., Yuan, W. and Ciais, P.: Improved global-scale predictions of soil carbon stocks with Millennial Version 2, Soil Biology and Biochemistry 164, 108466, https://doi.org/10.1016/j.soilbio.2021.108466, 2022.

Bailey, V. L., Bond-Lamberty, B., DeAngelis, K., Grandy, A. S., Hawkes, C. V., Heckman, K., Lajtha, K., Phillips, R. P., Sulman, B. N. and Todd-Brown, K. E.: Soil carbon cycling proxies: Understanding their critical role in predicting climate change feedbacks, Global change biology, 24, 895-905, https://doi.org/10.1111/gcb.13926, 2018.

Chen, B., Lu, Q., Wei, L., Fu, W., Wei, Z. and Tian, S.: Global predictions of topsoil organic carbon stocks under changing climate in the 21st century, Science of the Total Environment, 908, 168448, https://doi.org/10.1016/j.scitotenv.2023.168448, 2024.

Cotrufo, M. F., Ranalli, M. G., Haddix, M. L., Six, J. and Lugato, E.: Soil carbon storage informed by particulate and mineral-associated organic matter, Nature Geoscience, 12, 989-994, https://doi.org/10.1038/s41561-019-0484-6, 2019.

Duan, Q., Gupta, V. K. and Sorooshian, S.: Shuffled complex evolution approach for effective and efficient global minimization, Journal of optimization theory and applications, 76, 501-521, https://doi.org/10.1007/BF00939380, 1993.

Ghezzehei, T. A., Sulman, B., Arnold, C. L., Bogie, N. A. and Berhe, A. A.: On the role of soil water retention characteristic on aerobic microbial respiration, Biogeosciences, 16, 1187-1209, https://doi.org/10.5194/bg-16-1187-2019, 2019.

He, Y., Trumbore, S. E., Torn, M. S., Harden, J. W., Vaughn, L. J. S., Allison, S. D. and Randerson, J. T.: Radiocarbon constraints imply reduced carbon uptake by soils during the 21st century, Science, 353, 1419-1424, https://doi.org/10.1126/science.aad4273, 2016.

Jackson, R. B., Canadell, J., Ehleringer, J. R., Mooney, H. A., Sala, O. E. and Schulze, E. D.: A global analysis of root distribution for terrestrial biomes, Oecologia, 108, 389-411, https://doi.org/10.1007/BF00333714, 1996.

Koven, C. D., Riley, W. J., Subin, Z. M., Tang, J. Y., Torn, M. S., Collins, W. D., Bonan, G. B., Lawrence, D. M. and Swenson, S. C.: The effect of vertically resolved soil biogeochemistry and alternate soil C and N models on C dynamics of CLM4, Biogeosciences, 10, 7109-7131, https://doi.org/10.5194/bg-10-7109-2013, 2013.

Laub, M., Blagodatsky, S., Van de Broek, M., Schlichenmaier, S., Kunlanit, B., Six, J., Vityakon, P. and Cadisch, G.: SAMM version 1.0: a numerical model for microbial-mediated soil aggregate formation, Geoscientific Model Development, 17, 931-956, https://doi.org/10.5194/gmd-17-931-2024, 2024.

Lu, X., Wang, Y. P., Ziehn, T. and Dai, Y.: An efficient method for global parameter sensitivity analysis and its application to the Australian community land surface model (CABLE), Agriculture and forest meterorology, 182, 292-303, https://doi.org/10.1016/j.agrformet.2013.04.003, 2013.

Malone, B. P., Minasny, B. and McBratney, A. B.: Digital Soil Mapping. In: Using R for Digital Soil Mapping. Progress in Soil Science. Springer, Cham. https://doi.org/10.1007/978-3-319-44327-0_1, 2017.

Nyaupane, K., Mishra, U., Tao, F., Yeo, K., Riley, W. J., Hoffman, F. M. and Gautam, S.: Observational benchmarks inform representation of soil organic carbon dynamics in land surface models, Biogeosciences, 21, 5173-5183, https://doi.org/10.5194/bg-21-5173-2024, 2024.

Parfitt, R. L., Theng, B. K. G., Whitton, J. S. and Shepherd, T. G.: Effects of clay minerals and land use on organic matter pools, Geoderma, 75, 1-12, https://doi.org/10.1016/S0016-7061(96)00079-1, 1997.

Solly, E. F., Weber, V., Zimmermann, S., Walthert, L., Hagedorn, F. and Schmidt, M. W.: A critical evaluation of the relationship between the effective cation exchange capacity and soil organic carbon content in Swiss forest soils, Frontiers in Forests and Global Change, 3, 98, https://doi.org/10.3389/ffgc.2020.00098, 2020.

Wang, L., Abramowitz, G., Wang, Y.-P., Pitman, A. and Viscarra Rossel, R. A.: An ensemble estimate of Australian soil organic carbon using machine learning and process-based modelling, Soil, 10, 619-636, https://doi.org/10.5194/soil-10-619-2024, 2024.

Wieder, W., Grandy, A., Kallenbach, C., Taylor, P. and Bonan, G.: Representing life in the Earth system with soil microbial functional traits in the MIMICS model, Geoscientific Model Development, 8, 1789-1808, https://doi.org/10.5194/gmd-8-1789-2015, 2015.

Zhang, H., Goll, D. S., Wang, Y. P., Ciais, P., Wieder, W. R., Abramoff, R., Huang, Y., Guenet, B., Prescher, A. K. and Viscarra Rossel, R. A.: Microbial dynamics and soil physicochemical properties explain large-scale variations in soil organic carbon, Global change biology, 26, 2668-2685, https://doi.org/10.1111/gcb.14994, 2020.