The authors applied different models to predict the spatial variability of SOC concentration in different Australian ecosystems, using 1000 in-situ observations. The article is well written and structured and I think it should be considered for publication at EGUsphere. However, there are some flaws that should be addressed by the authors before publication.

We would like to thank the reviewers for their time and feedback on this manuscript. Please find our point-to-point responses below.

1. **First, the authors use SOC stocks to calculate SOC concentration using bulk density from the Soil and Landscape Grid National Soil Attributes Maps (SLGA). This is counter intuitive since SOC stocks cannot be measured but are usually estimated from SOC concentration and bulk density measurements. Why the authors don't use source data for SOC concentration? Was the SOC stock from their database initially calculated using the same bulk density data that they use?**

It's correct that we used bulk density from the SLGA data, however, our explanation wasn't clear, and we apologise for that. Let us explain. We did not use direct measurements of bulk density because they were not available. The recalculation of SOC concentration from SOC stocks and bulk density was to better compare the results from the simulations. We simulated SOC flux with MIMICS in mg/cm^3, and one of our datasets only had SOC stocks in t/ha. Unfortunately, the lack of new soil measurements over the country means that we can really only use datasets that are currently available. To compare the simulated SOC to the observations, we converted the units of both simulated SOC and observed SOC to g C/kg soil using the SLGA-derived bulk densities at each site. These unit conversions will not affect the results of MIMICS.

To compare the performance of MIMICS and machine learning models using metrics such as RMSE and MAE which are unit-specific, we also used SOC in g C/kg soil to train the machine learning models. We will clarify this in the revised manuscript.

2. **One of the main findings is that bulk density is one of the main predictors of SOC concentration. However, the common understanding is that SOC concentration is a predictor of bulk density so I think they should refine their modelling scheme to avoid biases in their result and revise the discussion on this part.**

Yes, that is correct. There is a clear (and known) negative relationship between SOC and bulk density. For example, higher bulk density usually corresponds to less porosity restricting plant root growth, oxygen availability to soil microbial communities and then SOC dynamics. Conversely, soil that is less dense has greater porosity, allows water infiltration, oxygen diffusion, root growth, microaggregate formation and enhances SOC dynamics. Therefore, it is reasonable to use bulk density as a predictor of SOC in our machine learning models. We'll revise the discussion to explicitly consider the relationship.

3. **Second, they use a vertically discretized version of MIMICS, but then only show results for the top 30 cm. What was the rationale behind using this vertically discretized model? I think it would be interesting to show the results – at least in the supplementary material – on the SOC concentration also for deeper layers. Even if they won't be able to compare it with observations, if possible, they can discuss their findings in comparison to the literature.**

Thanks for the suggestion. We used the vertically discretised version of MIMICS in this study mainly to consider the carbon loss from top 30 cm soil to deeper layers due to vertical transport processes. This aspect is lacking in the default version of MIMICS, where the only output of SOC is microbial respiration.

We compared the statistics between simulated SOC concentration in 30-150 cm soil by MIMICS (calculated as the mean value of 30-150 cm at 10 cm thickness) and SOC concentration extracted from a digital soil mapping product SLGA (calculated as weighted average of 30-60 cm, 60-100 cm and 100-200 cm soil). The mean value of SOC simulated by MIMICS-PFT (5.39 g C/kg soil) and MIMICS-ENV (6.54 g C/kg soil) are higher than that of SLGA (2.65 g C/kg soil) but lower than the mean value of the map representing 95% confidence interval of SLGA (6.99 g C/kg soil).

As noted by the reviewer, while we applied the vertically discretised version in this study, we only had observations for top 30 cm soil. The deeper layers were actually excluded from our parameter optimization process, they were not included in the calculation of the cost function. As SOC content in 0-30 cm is usually greater than that in deep soil layers, applying parameters optimized for 0-30 cm layer to deeper layers is likely to overestimate SOC at depth, making the predictions highly uncertain and unreliable. Therefore, we think it's less meaningful to present the results for deep layers in this paper. If the editor believes this to be important, we could compare statistics (e.g., min, max, mean value) of the MIMICS predictions of SOC in deeper layers to those extracted from the SLGA in a supplementary information file.

General comments

4. **Throughout the text, it is often unclear whether they refer to SOC stocks or SOC concentration. Please add this information each time when this is appropriate.**

Thanks, we'll clarify it in the revised manuscript.

5. **In the discussion, they should make an effort to better compare their results to the existing literature.**

Thanks, this was noted by Reviewer 1 as well, and as noted in our response there we'll improve our comparison with existing results and extend the discussion to include this.

6. **They chose to call one of their clustering groups plant functional types, but what they use are actually ecosystems. Please change this name throughout the text.**

Thanks for your suggestion regarding the classification of observations into ecosystems rather than plant functional types. However, we suggest that in this context "plant functional types" are appropriate, for two reasons,

1) The vegetation type is extracted from the National Vegetation Information Systems (NVIS) where the vegetation is divided into 32 "Major Vegetation Groups'. We aggregate them into four main types to guarantee sufficient observations for each group for model optimization. The aggregation is based solely on vegetation, without the consideration about other environmental characteristics.
2) "Plant functional types" are the classifications used in almost all process-based land models (e.g., land surface models) to represent plant and ecosystem diversity, so that using plant functional types here aligns better with existing literature and modelling approaches. This includes some existing applications of the MIMICS model (e.g. https://doi. org/10.1029/2020JG006205)

7. **The authors should add a map of the sites used and specify how many sites did they have for each ecosystem.**

Thanks for the suggestion. Figure 2 shows the spatial distribution and the number of sites in each plant functional type. We'll add a similar map for each environmental group in the revised manuscript.

Specific comments

8. **L23 I think plant functional types is not really a good definition of what you used. You used ecosystems rather than PFT.**

This is the same comment as 6 above. Please see our reply above.

9. **L28 when you say non-forest soils, you need to specify that you're not including croplands.**

Thanks, we'll clarify that in the manuscript revision.

10. **L45-46 reference needed.**

Thanks, we'll add references here (Jackson et al., 2017, https://doi.org/10.1146/annurev-ecolsys-112414-054234; Stockmann et al., 2013, https://doi.org/10.1016/j.agee.2012.10.001).

11. **L90-91 but also the parametrization and the lack of model validation with data (see Le Noe et al, 2023)**

Le Noë, Julia, Stefano Manzoni, Rose Abramoff, Tobias Bölscher, Elisa Bruni, Rémi Cardinael, Philippe Ciais, et al. "Soil Organic Carbon Models Need Independent Time-Series Validation for Reliable Prediction." Communications Earth & Environment 4, no. 1 (May 8, 2023): 158. https://doi.org/10.1038/s43247-023-00830-5.

Thanks for the addition. We'll revise it.

12. **L93 change "models has rarely" with "models have rarely"**

Thanks, we'll amend it.

13. **L91-94 Please add: and the difficulty of constraining the parameters of microbial explicit C models with relevant data on microbial properties, for example extracellular enzyme activities which are very hard to measure but also very important for these type of models.**

Thanks, we'll add it.

14. **L95 change "understanding on the" with "understanding of the"**

Thanks, we'll revise it.

15. **L96-98 You can also cite: Todd-Brown, K. E. O., J. T. Randerson, W. M. Post, F. M. Hoffman, C. Tarnocai, E. A. G. Schuur, and S. D. Allison. "Causes of Variation in Soil Carbon Simulations from CMIP5 Earth System Models and Comparison with**

Observations." *Biogeosciences* 10, no. 3 (March 13, 2013): 1717–36. https://doi.org/10.5194/bg-10-1717-2013.

Thanks, it's really a good paper and we'll cite it here.

16. L107-108 please specify what you mean with moderately well.

Thanks, we'll revise the sentence to make it clear.

17. L122 add "low" before amount of data.

Thanks, we'll add it this.

18. L129 add in forests, woodlands, shrublands and grasslands.

Thanks, we'll add that.

19. L143 change The" MIcrobial-MIneral Carbon Stabilization (MIMICS)" with "MIMICS"

Thanks, we'll revise it.

20. L144 change "soil carbon" with "SOC".

Thanks, we'll revise it.

21. L158 change "soil carbon" with "SOC".

Thanks, we'll revise it.

22. L180 average conditions during what period?

We used climate data from 1991 to 2020 in this study (Table 2).

23. L181 please specify what you mean by "a sensible range".

It means "a reasonable range" – we referred the papers but it was based more on experience to set the values because the pools are conceptual and cannot perfectly correspond to observed SOC fractions. We'll be more explicit about our choices here in the revised manuscript.

24. L194-196 "We trained the RF model with different numbers of trees": how many?

We set the number of trees as 100, 200, 300, 400 and 500, and the performance of random forest didn't vary much with different number of trees. We therefore chose a tree number of 200 in this study. We'll note this in the revised manuscript.

25. L203 use aggregate instead of segregate?

Thanks, we'll modify this.

26. L221 what predictors? please specify.

Thanks, we'll specify predictors here.

27. L223 Change permutation variable importance with PVI.

Thanks, we'll revise it.

28. L226 Change random forest with RF.

Thanks, we'll revise it.

29. L246 in Table 1, add the units of the parameters.

The parameters are scaling factors and are therefore unitless.

30. L250 change "The second approach was taking" with "The second approach consisted in taking".

Thanks, we'll amend this to 'the second approach used the most…'.

31. L251 specify which were the main predictors.

Thanks, we'll specify that.

32. L269-270 reference needed for SILO database.

We show the reference in Table 2. For better readability we'll move it to main text.

33. L275-278 In grasslands, or even in managed forests, NPP should be corrected by removing the fraction of NPP that is appropriated by human activities. Also, there might be high variabilities due to C input sources. The best would be to use different sources to see the uncertainties, but at least this should be discussed.

Thanks for the good suggestion here. We agree that the sources of NPP data will cause uncertainties in SOC estimations, and we'll add this part to discussion.

34. L281 reference needed for NVIS.

Sorry for the missing reference. We'll add it.

35. L285 reference needed for SLGA.

We show the reference in Table 2. For better readability we'll move it to main text.

36. L309-311 do these datasets use comparable methodologies for soil analysis?

The methods used to measure SOC concentration in the laboratory are different because they are from different projects, see Viscarra Rossel et al., (2019) and Bissett et al., (2016). However, the methods used to process and harmonise the datasets are the same and follow the approach in Viscarra Rossel et al., (2014). Briefly, these two datasets used are harmonised to represent SOC in the 0-30 cm layer using natural cubic splines for profiles with observation at more than 3 layers and a log-log model to profiles with observation at two layers. We will clarify that this is what we did.

37. L312 change "algorithm" with "equation".

Thanks, we'll revise it.

38. L317-318 specify the values used for S and I.

The equation is applied to each site, so the parameters S and I are site-specific.

39. L 332-336 move (a) to the beginning of the sentence. same with (b). Also, you need to add (a) and (b) in the figure.

Thanks, we'll revise it.

40. L340-342 Change the sentence as follows: For machine learning models, all observations were randomly separated into a training and test datasets, with […].

Thanks, we'll revise it.

41. L343 what groups?

Each plant functional type or environmental group divided using K-means. We'll clarify this in the revised manuscript.

42. L340-345 And then what values did you keep after cross-validation?

We selected the set of parameters which made MIMICS perform best on out-of-sample test data based on the metrics, please see L361-364 for details. And we'll add the optimized value of each parameter used to predict SOC at the continental scale in a new table.

43. L370 change "Permutation Variable Importance (PVI)" with PVI.

Thanks, we'll revise it.

44. L411 change random forest with RF.

Thanks, we'll revise it.

45. L415 add "but with a higher variability across sites" at the end of the sentence.

Thanks, we'll add it.

46. L455 change "approach" with "model".

Thanks, we'll revise it.

47. L457-459 but what about the relative standard deviation. is it still positively correlated with the mean? it's logic that higher mean ==> higher standard deviation in absolute terms, but it would be interesting to see if also the relative standard deviation is higher.

Thanks, it's really a good suggestion, we'll add the map of relative standard deviation.

48. L475 In table 2, add the results for the multi-model mean.

Thanks, we'll add that.

49. L483-484 add SOC spatial variations?

Thanks, we'll revise it.

50. L488 what does it entail in terms of uncertainty the fact that the number of predictors you used was lower than that employed in most digital mapping?

We think the uncertainty caused by the limited number of predictors is relatively small.

Most of studies that used machine learning for digital mapping of SOC typically employed different categories of predictors, including climate, organisms (e.g., plants), soil properties, terrain attributes (e.g., elevation) and land-use. In each category they usually selected several proxies, for example, temperature, precipitation, and evapotranspiration were widely used to represent climate conditions. Usually more than 20 predictors were employed in the digital mapping studies. However, some predictors contributed similarly to the predictability of SOC because of their collinearity, and some predictors contributed little because they were less important. Therefore, removing the redundant predictors and keeping only the most important ones (as assessed, for example, by random forest variable importance indicators) may actually provide not just comparable skill, but increased interpretability of SOC variations.

In our study, we used mean annual temperature, mean annual precipitation, soil clay content, bulk density, and NPP to train the machine learning model. These predictors were widely found to be important in SOC variations. Taking them as predictors, our random forest model showed good performance in predicting SOC concentration with R^2 greater than 0.9 for out-of-sample test data. In this case we think the predictors we used are sufficient to interpret SOC variations in Australia. While including additional predictors clearly may improve machine learning performance, the enhancement will not be large, given that R^2 is already greater than 0.9.

Equally important is limiting predictors to be comparable to the predictors given to the MIMICS model. This allows us to understand how well MIMICS is utilising the information available to it. Therefore, we opt to retain a limited number of predictors to facilitate a more meaningful comparison among different models.

51. L491 is there a bias because you calculated concentration from SOC stocks and bulk density?

We used SOC concentration (g C/kg soil) to train random forest model in order to compare models' performance based on metrics like RMSE and MAE (please see our reply to the first comment above). To test if the importance of bulk density on SOC concentration is universal, we

1) estimated variable importance using only forest SOC observations from BASE dataset where SOC is originally reported in concentration (%). Bulk density also ranked first among all predictors.
2) Estimated variable importance using a global SOC concentration dataset (WOSIS, https://www.isric.org/explore/wosis), and bulk density also ranked first in SOC concentration (unpublished result).
3) Referred to papers (e.g., https://doi.org/10.1007/s11104-015-2380-1) that also found the importance of bulk density on SOC concentration in Australia.

Therefore, we think the bias due to the calculation of SOC concentration using bulk density is small.

52. L493-494 And land-use?

Yes land-use significantly affects soil bulk density. We'll mention it here.

53. L501 change much more with much higher.

Thanks, we'll revise it.

54. L500-502 what other attributes did they use? does it mean that the number of predictors you used is not sufficient? Can you test that?

They used around 30 predictors relating to climate, soil properties and so on. The predictors we used are sufficient in this study, please see our reply to (50) above for details.

55. L517-518 add "in these ecosystems" at the end of the sentence.

Thanks, we'll revise it.

56. L521-523 due to what?

Soil bulk density in forest is much lower than that in soils under other vegetation types. This is a finding based on the analysis of observations. It may be attributed to the root penetration, the formation of aggregate and stable structure due to higher content of soil organic matters in forests. And organic matters have much lower bulk density than other soil constitutes such as clay, silt and sand, therefore, forest soils with high SOC content usually have lower bulk density than non-forest soils. On the other hand, lower bulk density improves the oxygen availability to microbial communities, accelerating their activities for carbon stabilization. We'll make the relationship between SOC and bulk density explicit in the revised manuscript.

57. L538 change process-based model with process-based models.

Thanks, we'll revise it.

58. L558 remove "two".

Thanks, we'll remove it.

59. L562 change moderately well with moderately good.

Thanks, we'll revise it.

60. L568 change "exit" with "exist".

Sorry for the typo here. We'll amend it.

61. L571 change "at 30 cm soil" with "at 30 cm depth".

Thanks, we'll revise it.

62. L573 add "were" between "values" and "used".

Thanks, we'll revise it.

63. L582-584 where does MIMICS get wrong?

Based on the findings in this study, we can first contribute the lower accuracy of MIMICS to the lack of explicit representation of a soil moisture effect. MAP is important for SOC variations especially in shrublands and woodlands (Figure 3) and taking it into account for parameter optimization (MIMICS-ENV) benefits the model performance (greater R^2 and LCCC in Figure 6) in these vegetation types. Though MAP clearly isn't the same as soil moisture, it highlights the importance of water in SOC dynamics (see L590-600 for details).

MIMICS is established based on our understanding on SOC turnover processes. However, some processes are simplified or ignored because of the lack of knowledge and observational data, as well as the increased uncertainties when new parameters are introduced. Machine learning models learn patterns from the data without relying on the explicit mechanisms of SOC dynamics, therefore can better capture and handle the complex relationships between predictors and SOC.

64. L586-588 change the sentence as follows: In their study, NPP and MAT had the highest explanatory power for SOC stocks according to MIMICS, while clay content had the highest explanatory power according to the SOC stock(???) observations, which limits the predictability of SOC using MIMICS.

Thanks, we'll revise the sentence here to make it clearer.

SOC is originally reported in concentration (%), but model output is in mg/cm^3. To compare model outputs and observations, we need to unify the unit using bulk density. When we say SOC stock observations, we mean converted observed SOC concentration to SOC stocks using bulk density to compare with model outputs. We'll make this clear in revised manuscript.

65. L588-590 Do you mean according to the random forest model applied to predict SOC concentration?

Yes, according to the PVI value showing the importance of predictors (Figure 3). PVI values are estimated using random forest.

66. L602-603 which means that...

SOC concentration in forest soils is much higher and more variable than that in soils under other vegetation types, which will likely lead to a higher MAE and RMSE. However, based on the higher R^2 and LCCC metrics for forest, we conclude that SOC in forest soil is more predictable because it aligns well with the observed patterns. We'll revise the discussion here to clarify this point.

67. L603-604 what do you mean? Do you mean it limits the accuracy?

The sentence here aims to explain why the MAE and RMSE for forest SOC is higher than those for other vegetation types (please see our reply to (66) above for details). Predictions of SOC in forest likely have larger absolute errors but also larger consistency to observed patterns. We'll rewrite this in the revised manuscript to make it clearer.

68. L606 add "that" after "Forests, given".

Thanks, we'll revise it.

69. L606-608 But then how comes that prediction error is systematically lower in non-forest ecosystems than in forests?

This comes down to the choice of metric used – they each give different information. If we consider the absolute error, prediction in non-forest is more accurate. But this is because the magnitudes are larger. A 10% error in a forest ecosystem might look huge compared to a 50% error in a semi-arid landscape. But if we consider the consistency and concordance between the pattern of predictions and observations, forest SOC is more predictable with higher R^2 and LCCC. We'll revise our discussion here to make this clearer in the revised manuscript.

70. L607 change "afford" with "show".

Thanks, we'll revise it.

71. L610-611 and a higher C influx due to animal residues?

Definitely. We'll add it.

72. L627 change random forest with RF.

Thanks, we'll revise it.

73. L628 change "more conservative" with "lower".

Thanks, we'll revise it.

74. L627-628 how much?

Sorry for the missing information, we'll add the citation of figure and table showing the results.

75. L673-675 Is there a relationship between C input and MAP in the data?

Yes, they're positively correlated with R^2 at 0.6.

76. **L691 why do the authors think that forest and grassland are explained by the same predictors, while shrubland and woodland by the same ones too? Unless this is already mentioned, please add something to the discussion.**

Sorry for the missing information. It’s concluded based on the analysis of predictor importance (Figure 3) but there’s a lack of text describing the results systematically. We’ll revise both the Conclusion and Result to make them more consistent.