

**Thank you for your replies to the Reviewer's Comments RC1 and RC2. I have some remarks regarding these replies, a number of own remarks and would ask to take these into account as well, while preparing the revised manuscript:**

We thank the Editor for the careful reading of our manuscript and for spotting many small mistakes. Below we answer to the main comments.

#### **Replies to RC1:**

**Most replies are convincing, with some room for improvement for these:**

**L73: convincing on novelty: the reply is clear, only be careful to use the words "model" and "processes" as it suggests a (semi-)mechanistic approach is taken, which is not truly the case. "models" has the meaning of "models of soil variation" I guess, but such models rather estimate the net effect, at sampling time, of various processes than the processes themselves. See also RC2 for that.**

**L80-92: The citation Jones, A. et al should be: Jones, A., Montanarella, L. and Jones, R. Soil atlas of Europe (...). First names and surnames were swapped.**

**L209 workflow: Perhaps a workflow figure is not needed, but a stepwise description would certainly help the reader. Now the focus is on technical details (parallelization, subsampling to ease computation: not evaluated so why named?). I suggest to rewrite 2.7 to illustrate the order of activities, of course with references to used packages, etc.**

We have added a stepwise description of the workflow at the beginning of Section 2.7. We have also checked the new reference to Jones et al., (2005) to not invert name and surname.

#### **Replies to RC2:**

**There is a substantial difference in opinion between Reviewer 2 and the Authors. In most cases I would follow the authors in their response.**

**Where I would follow reviewer 2 at least partly is here:**

**A. I do agree with Reviewer 2 that Shapely values could be introduced, in words, earlier in the manuscript (see comment about L71-72 below).**

**B. Figure 3 is indeed a bit difficult to interpret, see my own comment about L332, where I would arrive at the opposite conclusion as the authors.**

**C. About the soil processes versus the correlative relations to covariates. I think the authors do not pretend to "model" or even identify soil processes in a mechanistic way, but the reason for this could be better explained. They look at the "boundary inputs" (controls) to mechanistic models rather than the processes, translate these to covariates in an empirical study, and assume these boundary inputs were valid for the last decades to centuries (the time span during which SOC adapts to a -changing- environment including land use. See my comment on L413 below.**

We have now introduced Shapley values at the beginning of the last paragraph of the Introduction. We also checked the description of the results related to Fig. 3 and we found a small mistake, we meant  $T_{mwarmq}$  and not  $T_{mdq}$  in the first part of the sentence. We made the change, and the description is now correct. It does not change much the discussion because both are temperature variables. We have added more information regarding your comment at L. 413 of the original manuscript.

**Some additional comments:**

**L26: determinant>determinants**

We have made the change.

**L28: indicator>indicators**

We have made the change.

**L30: has>have**

We have made the change.

**L35: be>been**

We have made the change.

**L39: remains>remain**

We have made the change.

**L42: boosted regression> boosted regression trees**

We have made the change.

**L57: vary>varies (refers to importance)**

We have made the change.

**L71-72: It would be useful to already add here, in words rather than formulas, what Shapley values are. Section 2.6 is excellent but that is 100 lines further...**

We have added more information on the Shapley values as suggested.

**L89: Leptosol>Leptosols**

We have made the change.

**L95: If exactly the centre of the grid cell of 16x16 km<sup>2</sup> is sampled, it is not random sampling (representing the full grid cell) but systematic sampling. Perhaps explain better how a "sampling unit" at the grid centre still allows random sampling. I guess that a composite sample is taken at random positions inside the central 20x20 m<sup>2</sup> in each grid cell of 16x16 km<sup>2</sup> (unless inaccessible)?**

Systematic random sampling is not the same as simple random sampling, but it is random sampling design, see <https://dickbrus.github.io/SpatialSamplingwithR/SY.html>. Here, the grid is selected randomly over the area. It is a probability sampling design because it has the characteristics of a probability sampling design. Note that here our population is the whole of France, not the grid cell, so the objective of the sampling design is to provide an unbiased estimation of the population parameters for the whole of France.

**L98: The sentence suggests that alternative locations may not be cultivated, while in L100 it is stated to cover agricultural conditions. Perhaps reformulate.**

We have reformulated.

**L104: micrometer**

We have made the change.

**L109: mass or bulk density? Does this BD method include the stones (it seems so with the water bag method, if the stones are taken out of the pit).**

Please note that the next paragraph already mention that we work with mass.

**Table 1: annual mean solar radiations>annual mean solar radiation**

We have made the change.

**L140: based>based on**

We have made the change.

**L141: the a>a**

We have made the change.

**L143: user-define>user-defined**

We have made the change.

**L156: size approximately equal>approximately equal size**

We have made the change.

**L159-160: These are results, not methods.**

We agree that these are results. In this case, however, we would like to keep them there because in the Results section we only focus on the Shapley values, not on the model fitting. We prefer not to distract from the main message of the paper where modelling and mapping of SOC stocks is not the main result.

**L201: give>gives**

We have made the change.

**L207: require>requires**

We have made the change.

**L216: a feasible>as feasible**

We have made the change.

**L238: than>as**

We have made the change.

**L258: covariates>covariate**

We have made the change.

**L266: covariate>covariates**

We have made the change.

**L273: contribute>contributes**

We have made the change.

**L283: b)  $a > b$ ) an**

We have made the change.

**L294: MEC is not introduced, please explain**

We have made the change.

**L296: from covariate set>from the covariate set**

We have made the change.

**L301: is ME the same as MEC?**

We have now defined both with their acronyms earlier in the manuscript. ME is the mean error and MEC is the modelling efficiency, or Nash-Sutcliffe modelling efficiency.

**L331: ??>what figure?**

We have made the change.

**L332: The opposite looks true to me: In fig. 3, variations of SOC stocks induced by  $T_{am}$  appear to have a larger impact than of  $T_{mdq}$ . Please check.**

We have checked and made changes, see also our previous answer.

**L333: similar processes: could this be "similar but opposing processes"? Higher T = higher net primary production, but also higher decomposition rates. This explains why the over-all response of SOC to T is not linear.**

Thank you for the comment, we have added this information.

**L340: trend>a trend + contribute>contributes**

We have made the change.

**L351: you could mention the effects of erosion/deposition (i.e. redistribution) and exposition (affecting temperature and precipitation per unit area).**

We have added this information.

**L354: of SOC>on SOC**

We have made the change.

**L382: Do you have Podzols in Les Landes, and might these explain high SOC (in the subsoil, likely)?**

Yes the Landes have Podzols with accumulation horizons with high concentration of humic acids. We have added this information.

**L393: of SOC>on SOC**

We have made the change.

**L394: a single depth intervals>a single depth interval**

We have made the change.

**L395: factor>factors**

We have made the change.

**L413: It might be useful to add, that SOC is the resultant of the integrated net effect of processes over time, which for SOC can be decades to centuries. The covariates only relate to the current situation, which may be not representative (e.g. the climate-related and land use related covariates). This is why the predictive models may never reach high accuracies.**

We agree and we have added more information about this in the Discussion.

## **Review for “Shapley values reveal the drivers of soil organic carbon stocks prediction”**

**In this paper, Shapley values are used to interpret soil organic carbon variations nationwide in France using Shapley values. According to the results, this approach can explicitly explain the effect of soil-forming factors on SOC variation. This study was well-structured, well-written, and well-designed. There are, however, a few minor corrections and modifications that need to be made before publication, as listed below.**

We thank the reviewer for his/her positive evaluation of our manuscript. We address below all the comments and criticisms raised.

### **Abstract.**

**Currently, the abstract is mainly narrative, so some quantitative results would provide a better insight into the research.**

In the revised version we will add one quantitative result in the abstract.

**L30-31: Please provide citations to support the statement, “there has been studies that attempted...”**

We agree that it would be useful to add a few references to support this claim. In the revised manuscript we will provide some example references:

*Van Wesemael, B., et al. "Agricultural management explains historic changes in regional soil carbon stocks." Proceedings of the National Academy of Sciences 107.33 (2010): 14926-14930.*

*Wang, B., et al. "Modelling and mapping soil organic carbon stocks under future climate change in south-eastern Australia." Geoderma 405 (2022): 115442.*

*Rahman, N., et al. "Changes in soil organic carbon stocks after conversion from forest to oil palm plantations in Malaysian Borneo." Environmental Research Letters 13.10 (2018): 105001.*

**L64: remove “, Challenge 3” it is unnecessary.**

The paper on the ten challenges of pedometrics contains 10 challenges and we would like to keep citation to a specific challenge so as to be precise to as which challenge we are referring to. Many of the challenges from this paper are irrelevant to our study.

**L67: There was no explanation of the "SHAP" in this abbreviation.**

Here SHAP refers to the method from Lundberg et al. but we will add the definition of the acronym in the revised manuscript. SHAP stands for SHapley Additive exPlanations.

**L73: A key issue is convincing of the novelty of the research and highlighting the current research gap in the existing research. This is missing from the present manuscript.**

The main novelty of this study is to propose a method to interpret complex models used to predict spatially a soil property. This method relies on Shapley values, which has not been

described thoroughly in the soil science literature. To show the relevance of the use of Shapley value in interpreting a complex model, we use a study case in which the processes are well known and described (our French case study on SOC stocks), and use it to highlight that the Shapley values captured relationships between the SOC stocks and the environmental covariates that are meaningful. We believe the results presented in this study are novel and relevant to many studies mapping soil properties. For example, it is very common to report an estimate of the overall variable importance in prediction in soil mapping studies, but here we show that much more can be obtained, such as the partial dependence and the local importance. This is to our knowledge the first study in soil science showing how we can obtain a high level of insights into a complex model predicting a soil property.

**L80: “carbon stocks” replace with “SOC stocks” in case it is related to soil organic carbon.**

We agree and will make the change in the revised manuscript.

**L80-92: adding citations to this section is necessary.**

In the revised manuscript we will add the following two citations to this section:

*Laroche, B. et al. "Le programme inventaire gestion conservation des sols de France: volet référentiel régional pédologique." Étude et gestion des sols 21.1 (2014): 25-36.*

*Jones, A., Luca M. and Robert J.. Soil atlas of Europe. European Commission, 2005.*

**L89: Leptosols**

Thank you for spotting this mistake, we will make the change.

**L209: By adding a workflow of the study, the authors will make it easier for readers to follow the method's steps and understand the results.**

We understand but the manuscript is already quite long for a regular article (almost 8000 words and 9 figures), and the workflow is fairly simple and common to any mapping study: building of a regression matrix of soil properties with environmental covariates, fitting of a model, interpretation of the relationships found by the model (using Shapley values), and prediction.

**L224 “Fig. 2b further shows that the four most important covariates have”**

We will make the suggested change in the revised manuscript.

## **Shapley values reveal the drivers of soil organic carbon stocks prediction**

**This title is too methodological and provides no meaningful insights regarding what this study is reporting. I think the title will be meaningful if written as:**

### **Elevation, vegetation and temperature determine the spatial variation of French SOC stocks**

We disagree and do not see why our current title is too methodological. The study is effectively reporting on the Shapley values to determine the drivers of model prediction of carbon stocks.

The suggested title seems both misleading and inaccurate to us. On the three covariables that the reviewer suggested, only elevation is important. Vegetation is a group of covariable, not a single covariable. We have also several covariables related to temperature. Further, the relationship between the covariables and the SOC stocks is much more subtle than the fact that some variables “determine” SOC levels as suggested in the proposed title. The relationship changes with SOC stocks values, by landuse, and spatially. This is the purpose of our manuscript: to show that we can get much more than simply the average importance of variable. We would also like to avoid using the term “determine the spatial variation” because we do not have any mechanistic modelling involved in this study, we only determine the drivers of the model prediction. In other words, we determine which variables are important to the model, but we cannot say that these variables determine the spatial variation of the SOC stocks. We have a whole paragraph about this in the Discussion.

**Identifying relationships between environmental factors and SOC stocks is an important topic of scientific investigation. In this study, authors used soil samples from 2206 sampling sites and data of 23 environmental factors from France to predict the spatial variation of SOC stocks of 0-50 cm depth interval. Authors investigated how the correlations between SOC and environmental factors vary across the prediction points of the study area using “shapely values”. Authors reported that topography, reflectance property of vegetation (NDVI), and temperature primarily explain the spatial variation of French SOC stocks. I think authors are attempting to address an important topic, but this manuscript needs substantial revision before it can be published.**

**There has been a number of SOC stock studies previously published from France, which have reported relationships between environmental factors and SOC stocks. Authors should compare their findings with previous studies and explain how and why their result is different and novel. Authors should also report whether they used the soils samples used by previous studies, and which findings are new in this manuscript. To merit for publication, authors should explain what are new findings in this study that is not available in previous studies from the same study area.**

The reviewer may have missed the whole paragraph in Section 4.3 and called “Comparison with previous studies”. There are indeed several studies in France mapping SOC or SOC stocks. It is not the purpose of our manuscript to make another map of the SOC stocks. Instead, we want to show how the Shapley values are a useful methodological development to interpret a complex model. The fact that France has many studies on SOC stocks is very valuable in our case, because we use them to compare our findings and the relationships found by the models. We cite all these French case studies and we have a large discussion to discuss the relationship found by the model to existing studies for the same area.

**I am not comfortable in authors using “process-based” modeling phrase repeatedly in this manuscript. In this study, authors did not use any process-based model, nor they report any new soil carbon regulating process, so it’s just a pure distraction. There is a long and rich history of SOC process-based modeling literature where studies attempt to predict the temporal dynamics of SOC under changing land use and climate, which is not within the scope of this manuscript.**

We are surprised by this comment because we have clearly mentioned on several occasions in our manuscript that we should not infer causal mechanisms from correlation found in the data using empirical modelling. For example, at lines 406-408: *Despite that we selected a set of covariates that intended to represent underlying mechanisms involved in SOC storage, first, these are only proxy variables and do not necessarily relate to processes involved in SOC stocks variation.* The only mention of process-based modelling is in the introduction to explain the different possibilities to model SOC stocks spatially.

**In summary, I found this manuscript as prepared in rush, and does not report any interesting mathematical relationships between environmental factors and SOC stocks, which can be used to predict the SOC stocks. The manuscript is not focused and sentence structures need substantial revision before it can be published. My comments below are intended to improve the quality of this manuscript.**

It is difficult to understand the rationale for stating that our manuscript was prepared “in a rush”. We also do not understand why we should report mathematical relationships between SOC stocks and environmental variables: it is not the purpose of our study, and not the purpose of studies mapping soil properties over large areas. We are not fitting pedotransfer functions.

Saying that the manuscript is “not focused” and that “structures need substantial revision” without any specific comment is not really helpful. It is also not the opinion of the other reviewer who found the paper well-structured and well-written.

**Abstract:**

**I am not aware about the word limitation in the Abstract for this journal, but currently this abstract is more than 350 words and could be reduced substantially by deleting unnecessary texts. The abstract is not structured and should be rewritten. By reading the abstract, I couldn’t understand what was the relation between RF and shapely values, and why both are used in this study.**

It is unfortunate that the reviewer did not specify what is meant by “unnecessary texts”. The abstract is structured following the traditional way, with an introduction sentence, identification of the gap, proposed solution, methods, case study, results, and relevance of the findings. This is a widely accepted structure.

The relationship between Shapley values and RF is clearly stated: We introduce Shapley values, [...] and use them to understand how environmental factors influence SOC stocks prediction

**L4-11: These sentences describe the methods used in this study. Please replace these sentences and describe your methods briefly in 1-2 sentences.**

The sentences are not only about the method, but about the proposed solution, the test case and the approach. This is highly relevant in an abstract.

**L7: Please define what is shapely values, and why someone should care about it?**

On the one hand Reviewer wants less description of the method (previous comment) and on the other hand more description of the method. The current text is limited in length and description of the method is left in the Methods section. For the abstract, we believe that the current text is sufficient to understand what Shapley values are: *We introduce Shapley values, a method from coalitional game theory, and use them to understand how environmental factors influence SOC stocks prediction: what is the functional form of the association in the model between SOC stocks and environmental covariates, and how the covariate importance varies locally from one location to another and between carbon-landscape zones.*

**L8-9: “what is the .....”. This sentence is not correct. The relationships shown in Figure 3 are relationships between “shapely values and environmental factors”, and not the “relationships between environmental factors and SOC stocks”, which are not the same. Authors need to clarify this statement.**

We disagree, Figure 3 shows the relationship between The SOC stocks and the environmental variables. How to interpret Shapley values is described in the Methods section. Shapley values are expressed in the unit of the target variable. Figure 3 shows the partial dependence: how the SOC stocks vary for a change in the covariates.

**L10-12: In my understanding, this study reports correlational findings which may or may not be related to any soil carbon regulating processes, so I am not sure what “Results were validated both in light of the existing and well-described soil processes mediating soil carbon storage” means?**

This sentence means that we use the numerous studies available in France for mapping SOC stocks, to compare our correlation-based finding and interpretation results with past findings. We “validate” the relationships found in the model in light of the existing literature. We link the results of our studies with potential processes. This is explained in the first paragraph of Section 4.2: *The results suggest relationships between environmental covariates and SOC stocks which have been abundantly documented in the literature and other relationships that may highlight the limitations of empirical modelling for the SOC stocks prediction. Hereafter we describe how group of covariates relates to potential acting processes of soil carbon storage and how the Shapley values revealed potential limitations of the empirical modelling of SOC stocks.*

**L13-16: Again, these relations are based on correlations and does not provide any process-based understanding.**

We never claimed to derive new process-based understanding. On the contrary, we have put some warnings, please see Section 4.4.

**L16: “This shows...” I think this sentence does not report anything and not relevant in Abstract.**

We could remove this sentence in the revised manuscript.

## Introduction:

Introduction section should properly cite and discuss recent and relevant studies in this topic. I assume there are a number of studies which have attempted to explain the control of environmental factors on SOC stocks. Discussing the findings of these studies will strengthen this manuscript:

**Mishra et al. 2022. Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy as the machine learning, Soil Science Society of America Journal, doi:10.1002/saj2.20453.**

**Gautam et al. 2022. Climate change may release over 1.8 petagrams of soil organic carbon from topsoils in the United States by 2100, Global Ecology & Biogeography, 31, 1146-1160, doi: 10.1111/geb.13489**

We have cited many various studies in the Introduction, but we cannot/should not cite them all. This is not a review. Further our study is not about defining the control of SOC stocks, which has been done many times, but to show how Shapley values can interpret complex models of soil variation.

On the two references proposed above, the first could potentially be useful but the second seem irrelevant to our work.

**This is a spatial prediction study with no contribution to process-based modeling. So, the texts refereeing to process-based modeling is not relevant in this study and should be removed. I suggest discussing findings of additional studies which have reported mathematical relationships between environmental factors and SOC stocks.**

We consider the reference to process-based modelling here to be highly relevant for an Introduction. Process-based modelling is common to model SOC stocks, the purpose of the Introduction is to give some context. Why are we using empirical modelling when process-based models that do more justice to the well-known mechanisms of soil SOC storage are available? We need to provide reference to past studies on SOC stocks using process-based modelling to then highlight the need for our work. We never claimed that we do process-based modelling here.

**L35-36: "Dynamic modeling...". This sentence is not relevant to the content of this manuscript.**

See comment above.

## Materials and Methods

**Figure 2& 3: Please level the Y-axis in both figures, and provide units in both X and Y axis. Figure 3 does not provide any information regarding the relationships between environmental factors and SOC stocks, and I am not sure the scientific merit of these plots. Are these relationships additive, and can be used to predict the SOC stocks?**

We are not sure to understand this comment. Figure 3 shows the particle dependence: how does the SOC stocks values vary with changes in the covariates. Without further explanation

from the reviewer of why these plots have no scientific merit it is difficult to answer more precisely.

Indeed the relationships are additive. This is explained in Section 2.6, see the lines 189-190. To our knowledge the Shapley values are the only interpretation method available which enable additivity of the values. The sum of the Shapley values to the mean is the predicted SOC stocks.

**L 391: “We found .....”. This sentence suggests there were no new meaningful insights in this study.**

Previously, the reviewer criticized us for not comparing our results with previous studies, and now the reviewer says that our comparison suggests no new results. This is confusing. The reviewer may have missed that the purpose of our studies is NOT to map SOC stocks nor to get a better value of the validation statistics, but it is to show how the Shapley values can be used to interpret complex models. In our case, the fact that we found not notable difference with previous studies mapping SOC stocks in France is good news, the opposite would be worrying and we would need to investigate this further.

**L 444: Delete “Varied” from the sentence.**

We will make the change suggested.

**L450: Delete “full stop” from the middle of sentence.**

We will make the change suggested.

#### **References:**

**This is not a “literature review” manuscript, therefore, I encourage authors to give priority to recent literature of SOC stocks. For example, using studies published in the last 10 years in this topic unless the study is published from the same study area or have used the same set of samples.**

We disagree to give priority to recent papers. We even consider it bad practice. Why would one discard a publication if it is more than 10 years old? We include a reference because we think it is relevant, irrespective of the publication date.