EGUsphere, author comment AC2
https://doi.org/10.5194/egusphere-2022-1034-AC2, 2022


# Reply on RC2

Alexandre M. J.-C. Wadoux et al.

Author comment on "Shapley values reveal the drivers of soil organic carbon stock prediction" by Alexandre M. J.-C. Wadoux et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-1034-AC2, 2022

**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.