SOIL Discuss.,
https://doi.org/10.5194/soil-2020-75-RC2, 2021

Interactive comment on "Predicting the spatial distribution of soil organic carbon stock in Swedish forests using remotely sensed and site-specific variables" *by* Kpade O. L. Hounkpatin et al.

Madlene Nussbaum (Referee)

madlene.nussbaum@bfh.ch

Received and published: 14 January 2021

The manuscript shows the spatial prediction of carbon stocks in forests of Sweden for tow compartments (organic layer and mineral soil 0-50 cm depth) and the sum of the two. Random forest with parameter tuning and feature selection is used and prediction intervals are computed by quantile regression forest. Different model configuration is shown along two axes: 1) The data is split into subregions and separate model calibrations are presented. The results are then compared to a model covering the complete area. 2) Different subsets of covariates as model input are formed according

to their origin. Covariates only available at the observed sites and geodata available for whole Sweden were tested separately. In addition models combining all covariates were fitted and compared to the other two versions.

The methods are fully described and the results thoroughly discussed. The prediction intervals are validated with data not used to create the models which is so far rarely seen. Moreover, the distribution of the actual data is compared to the distribution of all predicted pixels. The manuscript is well written and the figures mostly are well prepared. Well done, Ozias!

However, I share the concerns of referee #1:

1. On page 6 lines 227–232 the data splitting approach is explained. I do not fully understand how it was actually done as I got puzzled by the last sentence. Were the validation data the same data points for the global and the local models (last sentence) or were they independently chosen (rest of this paragraph)?

   Nevertheless, the validation statistics should be computed on the exact same validation dataset. Especially the measure R^2 is very sensitive to different values of observed data. This means for the global model three local sets of validation statistics should be computed on the same data points. Of course overall statistics can also be presented, but should not be compared to the local statistics, because the standard deviation of the total vector of observed values might be larger.

   To achieve such a validation the 20-80% splitting should be done as stratified random sampling from the data points with the subareas as strata (which maybe already has been done).

2. The covariates only available at the observed locations cannot be used to create prediction at locations without any observation. Hence, they are not useful to

create area-wide maps unless one creates maps of these site specific characteristics beforehand. The same geodata would be used to create these predictions. Such a two-step prediction approach (some sort of a random forest co-kriging) is only meaningful if 1) there are more observations of this intermediate response (site characteristic) than for the final target response (SOC stock) which I understand is not the case for the current data set; 2) the intermediate response results for some reason in much better model performance and predictive accuracy.

The current study neglects to consider errors of the needed spatial prediction of the site characteristic covariates. The error of the spatial prediction might possibly be as high as to render these covariates useless. At least, the interpretation should put this clearly into perspective.

Moreover, please consider the small comments directly added to the manuscript using PDF annotations.

14.01.2021 / M. Nussbaum, BFH-HAFL

Please also note the supplement to this comment:
https://soil.copernicus.org/preprints/soil-2020-75/soil-2020-75-RC2-supplement.pdf

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2020-75, 2020.