SOIL Discuss.,
https://doi.org/10.5194/soil-2020-75-AC6, 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.

Kpade O. L. Hounkpatin et al.

ozias.hounkpatin@slu.se

Received and published: 1 March 2021

Dear Reviewer,

Please find enclosed the revision of our manuscript "Predicting the spatial distribution of soil organic carbon stock in Swedish forests using remotely sensed and site-specific covariates" by Hounkpatin et al. We thank you for the very competent review of our paper and the productive comments towards the improvement of our manuscript. We agree with almost all of them and revised the manuscript accordingly.

We hope that our paper is now acceptable for publication in SOIL.

Yours sincerely,

Ozias Hounkpatin

Answers to reviewer

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 conïňĄguration 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 ïňĄtted 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 ïňĄgures 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 R2 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 stratiïňĄed random sampling from the data points with the subareas as strata (which maybe already has been done). Author′s Response: Thanks for your enquiry. The error metrics were actually computed based on the same validation set for both global and local models, this to avoid comparison bias. Table 3 actually shows the error metrics for the global models when the independent validation set for the (different strata) North, Central and South Sweden were aggregated. Table 4 shows the specific error metrics for the local models against their respective validation set. Figure 2 presents the error metrics results of each global model for the three subareas along with corresponding local models (unfortunately repeating error metrics of local models from Table 4 for the independent validation. To ensure clarity, we provided more details in the section (see lines 224 – 231) related to the data splitting procedure. Manuscript lines (see lines 224 – 231): The RF models built on data covering the whole area of Sweden are hereafter called “global models”. The RF models created for each of the subareas are hereafter reported as “local models”. Considering the subareas as strata, the local models were built by randomly splitting the local datasets into calibration (80%) and validation (20%) subset. The training set of the global models was constituted by merging the 80% split of the strata dataset. To avoid comparison bias, both the global and local models were evaluated on the same validation set. Consequently, each global model was evaluated again three validation subset separately corresponding each to the 20% split of the Northern, Central and Southern local dataset. The merged 20% split of these local datasets was then used as validation set at national scale. We trained both global and local models on the calibration subset using tenfold cross validation with 5 repetitions using the R “caret” package (Kuhn, 2015).

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 speciïňĄc 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 ïňĄnal 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.

Author′s Response: We share the concern of the reviewer that all covariates should actually be maps for DSM of the whole area while the site specific data were only observations. Initially, the focus was mainly evaluating the geodata we now have versus site specific variables and not necessarily making a map. We made the map based on the geodata to give a visual trend of the distribution of the SOC stock. We also agree that the current study neglects to consider mapping uncertainties from potential maps of site specific variables which might actually translate into lower prediction accuracies of SOC stock compared to models based on their observation data. Our interpretation failed to put this clearly into perspective and we have now created a section discussing this issue. To elaborate more on this issue in that section, we made some preliminary maps (which need further improvements because of low kappa values) of the site specific variables using Random Forest and additional dataset from the inventory data. These maps were then used as covariates to predict SOC stock.

Manuscript lines (see lines 585 – 615): see section 4.4 Implication and limitations of the study DSM relies on existing maps for building regression models and ultimately prediction for mapping. The quality and accuracy of predictions depend as discussed earlier on choosing the most relevant covariates in relation to the target to be predicted.

The present study revealed that covariates which were available as maps did contribute to the MDA, but site characteristics were more prominent in relation to the SOC stock in Sweden. This might suggest that mapping these variables that were more decisive for SOC stock prediction and including them as covariates for mapping may improve accuracy. Since the primary focus of the present study was mainly to evaluate the GoC data versus the SSC data at different scale of modelling and not primarily for making a map, the observed data of the latter were used. However, since only the observed data of the site characteristics were considered, this study fails to consider that the mapping of these site characteristics would involve modelling errors and the propagation of these errors into the final maps of these site variables might actually reduce their predictive power. For completeness, we carried out a preliminary mapping of the site characteristics (SI-11) using additional soil inventory data, random forest and the GoC as predictors. These mapped site characteristics (mSSC) were then used as covariate for predicting the SOC stocks.

Table 6 presents the error metrics after independent validation for both local and global models along with the percentage margin of the RMSE in relation of the models based on GoC. First, the local mSSC based models still recorded lower RMSE as compared to global mSSC models. Compared to the GoC models, the overall positive percentage margin of the RMSE for the independent validations indicated that the mSSC models recorded the lowest RMSE. However, when assessing the RMSE margin between the global models of GoC and SSC, negative percentages were mainly recorded for the Northern Sweden independently of the depth. This indicated that the mSSC based global models were less accurate at predicting SOC stock locally in the Northern Sweden while the mSSC based local model did present a better prediction for Northern Sweden for the humus layer and mineral SOC stock. The mean SOC predictions based on the mSSC showed a stronger increasing gradient from Northern to Southern Sweden (SI-12) as compared to the pattern observed with the GoC maps. However, the uncertainty distribution was of similar magnitude (SI-12) as those observed for the maps based on GoC probably due to error propagations as these covariates were used

to make the site specific characteristics maps. This suggests that the SSC should still be supplemented for improvement at this stage with other covariates different from the GoC such as multi-temporal spectral (e.g. normalized difference vegetation index) data that are able to capture vegetation dynamic in forests. Notwithstanding the fact that was obviously error propagation, the study of which was beyond the scope of this study, the preceding results tend to confirm the potential of the high resolution maps of the site characteristics to contribute to the improvement of SOC stock prediction as compared to using only the GoC data. Given that the preliminary mappings of the SSC recorded low kappa (0.17 – 0.48) values (SI-11) at this stage further improvements are still necessary to improve their predictive ability and associated coefficient of variations.

Table 6: Cross-validation, independent validation of the global and local random forest models based on the mapped site specific covariates compared to models based on observed site specific covariates and grouped of covariates

Specific comments

Maybe consider using "covariates" instead of "variables" as in the rest of the text. It would be more specific. Author′s Response: “covariates” was considered Manuscript: Predicting the spatial distribution of soil organic carbon stock in Swedish forests using remotely sensed and site-specific covariates

Please add the reference depth of SOC stocks. Mineral soil = 0-50 cm (only), otherwise one would think down to bedrock or profile depth of usually 1-1.5 m

Author′s Response: We added the depth to the mineral soil. Manuscript (line 28 - 31): The most important covariates that influence the humus layer, mineral soil (0 – 50 cm) and total SOC stock were related to the site characteristic covariates and include the soil moisture class, vegetation type, soil type and soil texture.

Not all of the mentioned geodata are remote sensing data, e.g. the geological map. Better use a broader formulation. Author′s Response: We replace “remote sensing by

covariates" Manuscript (line 73 - 74): These studies and many others rely mostly on covariates existing as maps while survey data which present site specific information are left out during modelling.

Should "c." be "ca."? Author′s Response: yes it is ca. It has been replaced. Manuscript (line 99 - 100): Soil samples are collected in a subset of the plots with humus sampling on ca. 10 000 plots and mineral soil sampling on ca. 4500 plots (Stendahl et al., 2017).

This does not include the O horizon, right? Maybe reformulate "below the O horizon down to 30 cm". Author′s Response: Thanks for the suggestion. We reformulated to "below the O horizon down to 30 cm Manuscript (line 103 - 104): Therefore the content of inorganic carbon in mineral soil is considered negligible in the study area. Humus layer volumetric samples are taken using a soil core (core diameter 10 cm) below the O horizon down to 30 cm depth.

if available, better cite a data documentation that appears in the references list. Author′s Response: Citation has been added. Manuscript (line 120 - 122): Topographic covariates were computed from high-resolution digital elevation models (DEM) derived from Light Detection And Ranging (LiDAR) produced by the Swedish National Mapping Agency. It was originally created with 2-m spatial resolution (Dowling et al., 2013).

Please give the algorithm of the aggregation procedure. Bilinear interpolation? The software is usually secondary, if the algorithm becomes clear. Author′s Response: The bilinear interpolation algorithm was provided. Manuscript (line 122 - 124): However, the initial DEM was resampled in ArcGIS 10 software package using the aggregation procedure with bilinear interpolation to a final resolution of 10 m $\times$ 10 m which is reasonable for the data considered in the present study.

Please cite, or give used algorithms and versions. (maybe add in covariate table). Author′s Response: Citation has been added. Manuscript (line 124 - 125): The topographical covariates were computed using the SAGA GIS software (Conrad et al.,

2015).

Please cite in reference list as a web source. Author′s Response: Citation added. Manuscript (line 127 - 128): The depth to groundwater was obtained from the Swedish Forest Agency (SGU, 2018) and computes the difference in elevation in relation to surrounding cells following the vertical ïňĆow path.

Please add citation here. Or was this done in this study? Author′s Response: This was done in this study Manuscript (line 141 - 143): The resulting gamma-ray data as well as the geochemical data were interpolated in this study into maps either by ordinary kriging or inverse distance weighing when geostatistic assumption such as normal distribution were not met.

Please check again this function notation. The superscript 1 is likely an indicator function, usually used without superscript. Moreover, "j" is not defined in the text. Author′s Response: Thanks for the observation. We have checked and reproduce it as mentioned in Meinshausen (2006). j is also defined. Manuscript (line 182 - 185): The weight vector (Meinshausen, 2006) w_i is defined as follows: w_i (x,θ)= 1_{x_i Ïţ R_(l(x,θ)) } /(#{j:x_j Ïţ R_(l(x,θ)) }) (2)

Please indicate which values where tested for mtry in the grid search. Was ist 1:p with p: number of covariates? Author′s Response: We provided the values tested: 2: n, with n the number of covariates. Manuscript (line 196 - 198): To reduce computational load, the ntree was set at 500 while the mtry was tuned using the grid search (2: p, with p the number of covariates) method in the R "caret" package (Kuhn, 2015).

Please give a citation. e.g. Hastie et al., 2011. Author′s Response: Ciatation added. Manuscript (line 199 - 198): The importance of each input predictor can be assessed by the RF based on the mean decrease accuracy (MDA) (Hastie et al., 2011).

Were 10 m overlap really enough to allow for a smooth transition? This is just one pixel... Author′s Response: Thanks for notification. This was a mistake. We corrected

to 4km which was actually used. Manuscript (line 208 - 210): A buffer of 4 km was considered for the shapefiles of each subarea to create overlapping zones which ensured smooth transition while merging by averaging the SOC stock values within these shared units.

Please specify what kind of algorithm and maybe also the function name you used for de-correlation. Author′s Response: The function has been added. We use the Pearson′s correlation. Manuscript (line 220 - 222): Moreover, to reduce computation time while keeping relatively the same level of accuracy, we (1) used feature preprocessing capabilities implemented in the caret package (Kuhn, 2017) of R to remove highly correlated (Pearson's correlation) expressions using a cutoff point of 0.80. . .

maybe use "group of covariates" Author′s Response: Thanks for the suggestion. We adopted “group of covariates” throughout the document.

Please specify: How was the splitting done into 20 vs. 80%? Simple random sample? Author′s Response: Yes a simple random split was carried out. This has now been specified. Manuscript (line 229 - 230): Considering the subareas as strata, the local models were built by randomly splitting the local datasets into calibration (80%) and validation (20%) subset.

Bias is reported in the table, but definition is missing. Author′s Response: Thanks for notifying. Bias has now been defined. Manuscript (line 240 - 245): Bias= μ_pred-μ_obs

squared correlation coefficient, but rather the formula of the type: R2 = 1 - sum(P_i - O_i) / sum(O_i - mean(O)). Caret has an implementation, see also help site for the formula: R2(p,o, form = "traditional") Author′s Response: Thanks for notifying. This has now been corrected. Manuscript (line 240 - 245): Rˆ2 =1- ($\sum$_(i = 1)Θn(P_i − O_i)Θ2)/($\sum$_(i = 1)Θn(O_i−μ_obs)ˆ2) upper end, not at both ends simultaneously. But, the one-sided version of this plot is harder to read. However, the lines in the plots follow the 1:1-line quite well, so little difference is to be expected. Author′s

Response: In our study, the lines follow the 1:1 line quite well and we also thought that other upend end make no difference in this case.

Just to be precise ;-). Please consider using the peer-reviewed paper of this work. Nussbaum et al. 2014 (https://gmd.copernicus.org/articles/7/1197/2014/). We used an external-drift kriging model (that includes a linear regression of course). Then, the data was not only form the Swiss National Forest Inventory. Considerdropping this information as it is of no importance to the discussion. Author′s Response: We have removed the reference.

A bit hard to read. Consider reformulation. Author′s Response: We have reformulated that section and hope that it is now better to read. Manuscript (line 419 - 430): Low explained variances in predictive modelling could be related to different factors (Nelson et al., 2011). For example, the omission of key covariates with greater explanatory power or conversely using non-essential covariates with very low explanatory power which only increase the prediction error variance. Omitting key covariates in relation to SOC stock for forest ecosystem in the present study is less likely since covariates considered in this study well represent the surrogates for soil forming factors considered in the SCORPAN equation defined by McBratney et al. (2003). In addition, the removal of redundant and non-informative covariates was carried out via dimension reduction with the exclusion of highly correlated covariates and elimination of some others via recursive feature elimination. However, the Pearson correlations (min = 0, max = 0.28) between covariates and the different SOC stock were found to be poor though significant for most of the predictors (Table 5, SI 2-4). This could be expected because the data cover a wide range of different site conditions, soil types and parent materials.

Figure SI5 Author′s Response: Correction made. Manuscript (line 436): On the other hand, applying geostatistical approaches (SI-5) for the humus layer. . .

Computing variograms from the response directly might reveal spatial autocorrelation. However, I do not recommend to do that as putting as much of the correlation into

the regression part e.g. with random forest is to be preferred. Better reformulate the text. Author′s Response: We have reformulated focusing on the response directly. Manuscript (line 435 - 438): On the other hand, applying geostatistical approaches (SI-5) for the humus layer, mineral soil and total SOC stock revealed very low spatial autocorrelation for the different SOC stock suggesting that the structure of these SOC data is having a shorter range than the sampling interval.

Is it not the other way around, overestimation of small values and underestimation of the large values? RF has the tendency to predict the mean if correlation of response and covariate is weak. Hence, the extremes are not well predicted. Author′s Response: Yes generally there is overestimation of small values and underestimation of the large values. We were commenting on the trend which can be seen in Fig. 3. We mainly observed an underestimation of the low and high values in our context.

Conclusion difficult, without considering the error of such predicted maps. Author′s Response: We now provided error metrics based from RF models based on preliminary mapping of the site specific covariates which showed that they performed generally better than the group of covariates. Manuscript (line 580 - 607): Section on 4.4 Implication and limitations of the study DSM relies on existing maps for building regression models and ultimately prediction for mapping. The quality and accuracy of predictions depend as discussed earlier on choosing the most relevant covariates in relation to the target to be predicted. The present study revealed that covariates which were available as maps did contribute to the MDA, but site characteristics were more prominent in relation to the SOC stock in Sweden. This might suggest that mapping these variables that were more decisive for SOC stock prediction and including them as covariates for mapping may improve accuracy. Since the primary focus of the present study was mainly to evaluate the GoC data versus the SSC data at different scale of modelling and not primarily for making a map, the observed data of the latter were used. However, since only the observed data of the site characteristics were considered, this study fails to consider that the mapping of these site characteristics would involve modelling

errors and the propagation of these errors into the final maps of these site variables might actually reduce their predictive power. For completeness, we carried out a preliminary mapping of the site characteristics (SI 11) using additional soil inventory data, random forest and the GoC as predictors. These mapped site characteristics (mSSC) were then used as covariate for predicting the SOC stocks. Table 6 presents the error metrics after independent validation for both local and global models along with the percentage margin of the RMSE in relation of the models based on GoC. First, the local mSSC based models still recorded lower RMSE as compared to global mSSC models. Compared to the GoC models, the overall positive percentage margin of the RMSE for the independent validations indicated that the mSSC models recorded the lowest RMSE. However, when assessing the RMSE margin between the global models of GoC and SSC, negative percentages were mainly recorded for the Northern Sweden independently of the depth. This indicated that the mSSC based global models were less accurate at predicting SOC stock locally in the Northern Sweden while the mSSC based local model did present a better prediction for Northern Sweden for the humus layer and mineral SOC stock. Notwithstanding the fact that there could be error propagation, the study of which was beyond the scope of this study, the preceding results tend to confirm the potential of the high resolution maps of the site characteristics to contribute to the improvement of SOC stock prediction as compared to using only the GoC data. Given that the preliminary mappings of the SSC recorded low kappa values (SI11) at this stage further improvements are still necessary to improve their predictive ability. maybe use "open source" Author′s Response: Thanks for suggesting. “open source” adopted. Manuscript (line 579 - 583): DSM approach in the present study allows: flexibility in future improvement upon acquisition of new covariates or data point, repeatability in modelling with the application of the same modelling principles using open source software (e.g. R) and capitalizing on multi-source information (topography, site characteristics, forest data, gamma radiometry and geochemical data).

In the correlation matrix in the supplement (Figure SI2-4) the aspect is given. As it is 0-360 degrees it is not meaningful as a covariate. Consider dropping it everywhere.

Author′s Response: We dropped aspect in mapping the site covariates. In most cases, it was also dropped by RFE when all the variables were considered.

Please check again. These skewness values seem not possible from the other information. Author′s Response: We are sorry for this mistake. The n values were repeated twice. We have now corrected this in the table.

What kind of test was performed? To be complete it would be good practice to report the test statistic Author′s Response: Thanks for pointing out this. We have now changed the title of the table which take into account the test. Manuscript: Table 5: Random Forest variable importance for the global and local models for the humus layer, mineral soil and total SOC stock with their associated Pearson′s coefficient of correlation with the covariates (values in bracket, *: $p \leq 0.05$, $p \leq 0.05$;**: $p \leq 0.01$;***: $p \leq 0.001$)

Please specify how the R2 and the RMSE were computed. Cross-validation independent data set obtained by data splitting? Author′s Response: We have now specified that the r2 and RMSE were based on the independent data. Manuscript (line 950): Figure 2: Local and global models ranked by decreasing RMSE per subareas and category of covariates along with corresponding R2 based on the independent validation dataset. A: litter layer, B: mineral soil layer, C: total soil layer, SSC: site specific covariates, GoC: Group of covariates, allC: all covariates

These maps are very difficult to read. Consider either adequate generalization of the 10 m pixels for the display scale or showing but a section. I would actually be interested if there are artefacts at the border of the subareas like sudden change in values that are hard to explain. Such artefacts might be a reason against using local models even if their model performance is better. Author′s Response: Thanks for pointing this out. It is actually challenging to have a color palette that display distinctly the pattern over the whole Sweden. We have now used another color palette for the SOC stocks and hope that visualization has improved. Visualizing further the maps,

we did not observe any sudden change in values that would raise particular attention as the pattern at the border showed similar pattern for some location inside Sweden.

Please also note the supplement to this comment:
https://soil.copernicus.org/preprints/soil-2020-75/soil-2020-75-AC6-supplement.pdf

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