

Interactive comment on “High-resolution digital mapping of soil organic carbon in permafrost terrain using machine-learning: A case study in a sub-Arctic peatland environment” by Matthias B. Siewert

Anonymous Referee #4

Received and published: 3 October 2017

Author present comparison of four digital soil mapping techniques in predicting high-resolution (2x2m) SOC stocks of sub-Arctic peatland terrain. Study reports that Random forest performed better in comparison to other three techniques used and land cover types derived from a high resolution remote sensing data was the most important predictor of SOC stock variability. Author also report that most of the SOC of study area is relatively new carbon (~2000 years old).

Author report interesting findings and the outcome should be of interest to a wide readership of Biogeosciences. However, the current manuscript can be improved in multiple

C1

different ways as suggested below:

- The sentence structure at multiple places is awkward so a careful editing is required.
- Its not clear to me how 2x2m spatial resolution for SOC stock was defined? Author seem to have a variety of environmental datasets with spatial resolution ranging from 1 m to 20 m. - I don't agree with the term internal validation used in this manuscript. Using model training dataset as a model validation is not correct. It provides an incorrect metric of map accuracy. For validation, you have to either use the split sample in the beginning (like you did for 20% data) or it has to be take one out approach (cross validation; using remaining samples to predict at the data point by taking out that data point from the model calibration data).
- Its not clear to me how land cover data was treated in different models used, were all the land cover types were equally important predictors of SOC? or it was only a subset of all the land cover types? Please provide results.
- I will like to see a section on uncertainty in this manuscript. Either calculate the uncertainty or provide a discussion of potential sources of uncertainty involved this study.
- The manuscript will benefit if authors can provide reasoning to the observed results. For e.g., why the environmental predictors changed with depths, why certain environmental controllers were significant predictor at certain depth and not other.
- How the multicollinearity and non-linear relationships were handled?
- Fig. 5 need to be replaced, please remove pseudo sampling points from the plots, provide the number of samples used for model validation. Provide separate plots for 4 mapping techniques using validation samples only. Add R2, RMSE, and CCC values in each plots.
- Table 1: Please remove metrics calculated using model calibration datasets, and after adding these values in plots suggested earlier, you will not need this table. In results section, please describe what readers should learn from these map accuracy measures.

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2017-323, 2017.

C2