Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-552-RC1, 2018

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on "Mapping soil hydraulic properties using random forest based pedotransfer functions and geostatistics" *by* Brigitta Tóth et al.

Anonymous Referee #1

Received and published: 30 November 2018

Reviewer comments on "Mapping soil hydraulic properties using random forest based pedotransfer functions“ by Brigitta Tóth et al.

Summary In the manuscript by Brigitta Tóth et al. maps of soil water retention characteristics (i.e. soil water contents at saturation, field capacity and the wilting point) are derived for the catchment of lake Balaton from direct measurements (MARTHA data) and additional spatial information on soils, vegetation, topography and climate. In a first step the applicability of two tree-based machine learning algorithms was tested with the result that random forest outperformed generalized boosted regression models. In a second step random forests were combined with classical geostatistical methods to

predict the soil water retention characteristics. However, in most cases the combination of both approaches did not improve the predictions. Resulting maps and pedotransfer functions will be published for non-commercial use.

General comments The study presented in the manuscript is interesting and relevant since spatial information about soil water retention characteristics at regional scale is required for various purposes (e.g. as input data for regional hydrological models or crop modelling). The methods used to predict soil retention characteristics are adequate. However, the procedure of variable selection does not become totally clear. The manuscript is largely well structured, minor changes are suggested in the specific comments below. The conclusion is not in an appropriate form at all and should be written again. I followed the link in the manuscript but could not download the maps and pedotransfer functions. My overall impression is that the work deserves to be published in HESS after major revisions.

Specific comments Abstract P1 L16-17: Please formulate more precise: “water content at saturation (THS), at field capacity (FC), and at the wilting point (WP)”

Introduction P2 L29 – P3 L2: In this paragraph only studies are listed in which tree-based MLA algorithms worked best. Are there also studies where other methods like e.g. artificial neural networks performed best? If yes, they should also be mentioned here. I also think that tree-based methods are a very good choice in this study, but I wonder if there is really only one best approach. P3 L10: Do you mean soil water content at field capacity and wilting point? P3 L11: How can measurements be optimized? What is meant by number of measurements? A large number of? P3 L14-21: Please provide some numbers summarizing the uncertainties found in the studies cited, so the reader can get a feeling about which order of magnitude of uncertainties can be expected when predicting soil retention data. This might also define the “internationally accepted performance of hydraulic PTFs” mentioned in the abstract (P1 L22). P3 L23-25: The objective of the study should be clear and unambiguous. The formulation of the aim(s) should therefore always be identical when mentioned in the text (in the abstract, in the last paragraph of the introduction and in the first paragraph of the conclusion).

Materials and Methods P4 L12-20: The quite large number of abbreviations introduced in the manuscript unnecessarily demands the capacity of the reader. Please omit abbreviations when the term is used only a few times (e.g. ST or PSD). P4 L22: The covariates are only used to predict the soil hydraulic properties. The relationships between the response and predictor variables are not analysed in the manuscript (e.g. by partial dependence plots). Please rephrase “analysis of the relationships”. P4 L22: What does the number 173 stand for? Is it the number of available covariates? P4 L30 – P5 L2: I had to read the sentences several times to understand them. Please rephrase. P4 L25 – P5 L5: The content of the paragraph is not really covered by the heading “Soil hydraulic dataset”. Please adapt the heading. I also asked myself, if some information should be shifted to section 2.4.1. P5 L7: Please rephrase “most often used soil water retention values”. P5 L7-8: Why did you map water content at -330 cm matric potential when field capacity is determined in Hungary at -300 cm? P5 L15: Why are these methods the most efficient MLAs? This is a very general statement. I am sure that many data scientist would at least partially disagree. Please rephrase. See also my comment on P2 L29 – P3 L2. P5 L17: To calculate quantiles during the predictions? Quantiles of what? What is meant by “during the predictions”? P5 L15-L19: Please add some general information about the principles of regression trees. Also an unexperienced reader should get at least an imagination how the input information is transformed to water retention characteristics. Please also mention once the alternative names of the MLA′s (e.g. boosted regression trees) to avoid confusion. P5 L19: . . .build ensembles of models. . . P5 L19: . . . the difference between GBM and RF is the way. . . P5 L26: mtry? Seems to be an argument of an R function? P6 L1: 50 independent variables out of how many? I assume that it is related to the number 173 in P4 L21. Right? P6 L1: It is not really clear to me how you performed the variable selection. Especially when potential predictors are correlated it can be quite challenging to find an optimal set of predictors. Did you start with all possible predictors at once

or did you try out many different combinations of predictors? P6 L8-11: Terminology again: Is it right that “out of bag sampling” is identical to “bootstrapping”? If yes, you might also drop the term “bootstrapping” once. P6 L18: . . .to the median and the 5% and 95% quantiles. . . P6 L20 – P7 L8: The combination of state of the art MLAs and classical geostatistical tools seems plausible and promising to me. However, I wonder if it is correct to call it simply “direct mapping”. Isn′t it a combination of both: indirect (prediction with RF) and direct (kriging) mapping? Maybe I just haven′t understood the essential differences between direct and indirect mapping approaches. P6 L26: . . . Table 2 summarizes the measured. . . P7 L2: Here it says “most important covariates” (the result of the variable selection, right?), but in the caption of Table 1 it says (all) “available environmental covariates”. P7 L11: . . .with the method. . . P7 L10: “. . .based on measured soil hydraulic properties calculated for. . .”. How can the measured properties be calculated? Please rephrase the sentence.

Results and discussion P7 L27-28: In P6 L1 it says that most important 50 independent variables have been selected. How did you select them out of the 69-76 and 65-77 variables mentioned here? P8 L9-16: This paragraph should be shifted to the Materials and Methods section. P8 L6: Why can you assume that multicollinearities are no problem at all? I assume, that many of the predictors presented in Table 1 are highly correlated. I wonder if it is even possible to estimate a unique set of regression-tree parameters when predictors are correlated. For the same reason I could also imagine that it is not possible to determine one unique set of 50 most important independent variables. P8 L22: Please compare the values listed in the text and in Table 1 once again. I am not sure if they match. P8 L27 and many other passages in the text: Is it correct to use the term “covariate” when talking about regression trees? To me “predictors” or “independent variables” seems more plausible. P8 L30: . . .than soil related variables. . . P9 L6-L26: Please explain in the Method section how relative importance is determined. P9 L30: mtry? See also my comment on P5 L26. P9 L27 - P10 L13: In addition to the quality criterions presented in Table 4 it would be interesting to see scatterplots (measured versus predicted values). They sometimes give a better feeling for model performance and they also show if there are areas in the predicted data space of THS, FC and WP with very good or poor prediction performance. P10 L6-13: Please discuss what it is good for to add kriged values computed with a pure nugget model when the residuals of the RF predictions show no spatial structure. This way you simply add random numbers that blur your predicted mean values. I wonder, if you should leave out the whole exercise. P10 L10: the correlation is based on only three pairs of values. Please use a weaker formulation. P10 L33 – P11 L9 and Fig. 5: Why did you select WP in Fig 5 and why did you only show confidence intervals for HUN-PTF? It would also be interesting to see maps of THS and FK and the confidence interval from the RFK predictions. P11 L12: . . .we have not differentiated uncertainty of. . .

Conclusion P12 L1 – P13 L3: The conclusion has poor quality and should be written again. A conclusion should just consist of one or two paragraphs where the most important results are summarized and the most important conclusions are drawn. A concise take home massage can be formulated. In the following just some examples of aspects are listed that are wrong placed the conclusion of the manuscript: P12 L17-20: Such general methodological aspects are not the take home message of the study. P12 L30-32: A discussion of methods or suggestions of alternative methods should be done in the discussion section. P12 L31 – P13 L2: The conclusion is the wrong place for such a detailed discussion of the methods used. A new table (Table 7) should not be introduced in the conclusion section. P13L2-3: A comparison with findings by other authors should be done in the discussion section. New references should not be introduced in the conclusion (e.g. Webster and Oliver (2017)).

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-552, 2018.