Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2018-552-AC1, 2019

Hydrology and Earth System Sciences
Discussions

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

Brigitta Tóth et al.

szatmari@rissac.hu

Received and published: 11 February 2019

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.

A: Thank you for the detailed review and suggestions which help us to improve our manuscript. We hope to address the comments in a revised version of the article. Below we would like to answer and provide possible solutions for the comments and recommendations, following the referee's questions.

General comments:

Q1. The methods used to predict soil retention characteristics are adequate. However, the procedure of variable selection does not become totally clear.

A: With the variable selection our aim was to exclude less important predictors. Please find detailed description on how variables have been selected under answers for the specific comments.

Q2: 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.

A: Thank you for calling our attention to reformat the conclusions. Information will be moved under results and discussion and conclusions will be rewritten. Please find more information on it under answers for specific comments (P12 L1 – P13 L3).

Q3: I followed the link in the manuscript but could not download the maps and pedotransfer functions.

A: We are sorry that the download link of the maps did not work, something happened with the access authorization after submitting the manuscript, now the problem has been fixed.

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)"

A: Thank you for the suggestion, information on matric potential values will be added in the abstract as well.

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.

A: Thank you for your comment. We will provide some papers, which used several MLAs (e.g. neural networks, cubist, gradient boosting) for mapping soils and short summary of them: "Adhikari et al. (2014) used cubist combined with kriging for mapping soil organic carbon concentration and stock in Denmark and they found that cubist was appropriate for this purpose. The same was observed by Matos-Moreira et al. (2017), they used cubist for mapping the phosphorus concentration in north-western France. Behrens et al. (2018) compared a number of state of the art digital soil mapping methods including geostatistical techniques (i.e. ordinary kriging, regression kriging and geographically weighted regression) and machine learning algorithms (i.e. multivariate adaptive regression splines, radial basis function support vector machines, cubist, random forest and neural networks). They obtained the best results with cubist, random forest and bagged multivariate adaptive regression splines."

P3 L10: Do you mean soil water content at field capacity and wilting point?

A: Thank you for highlighting it, yes.

P3 L11: How can measurements be optimized? What is meant by number of measurements? A large number of?

A: You are right, the sampling density was optimized, the sentence will be clarified accordingly.

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).

A: Indeed, it is important, thank you for the idea. In predicting soil water retention data root mean squared error between 0.02 and 0.07 cm3 cm-3 can be expected depending

on the predicted soil hydraulic property and available input information (Nguyen et al., 2017; Zhang, Schaap, 2017; Román Dobraco et al., 2019). When PTFs are applied for mapping then uncertainty of the input soil layers will further increase the uncertainty of the PTFs, e.g. in point based validation RMSE was 0.073 cm cm-3 for FC mapped for China in Wu et al. (2018); Leenars et al. (2018) found that mean RMSE for THS, FC and WP together was 0.102 cm3 cm-3 for African soils; in EU-SoilHydroGrids RMSE was 0.095, 0.096, 0.084 cm cm-3 for THS, FC and WP respectively for European samples (Tóth et al., 2017). We will add this information in the text with the following references.

Leenaars, J. G. B., Claessens, L., Heuvelink, G. B. M., Hengl, T., Ruiperez González, M., van Bussel, L. G. J., Guilpart, N., Yang, H. and Cassman, K. G.: Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa, Geoderma, 324(February), 18–36, doi:10.1016/j.geoderma.2018.02.046, 2018.
Nguyen, P. M., Haghverdi, A., de Pue, J., Botula, Y.-D., Le, K. V., Waegeman, W. and Cornelis, W. M.: Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils, Biosyst. Eng., 153, 12–27, doi:10.1016/j.biosystemseng.2016.10.013, 2017.

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).

A: Thank you for the suggestion. We will rephrase the objective mentioned in the text to be identical. We will stick to the following main aim: to analyse difference in performance and spatial patterns between soil hydraulic maps derived with indirect (using PTFs) and direct (geostatistical) mapping methods. The possibility for a non-computation intensive method to map uncertainty of calculated soil hydraulic parameters is a possible advantage of the PTF method.

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).

A: We will decrease the number of abbreviations, we will not use MLA, OM, CaCO3 and EU-SHG in the text and will also remove unnecessary ST and PSD abbreviations. In the abstract we will not introduce the THS, FC, WP and MARTHA abbreviations.

P4 L22: The covariates are only used to predict the soil hydraulic properties. The relationships be- tween the response and predictor variables are not analysed in the manuscript (e.g. by partial dependence plots). Please rephrase “analysis of the relationships”.

A: Thank you for highlighting it, the sentence will be rephrased: “For the prediction of soil hydraulic properties based on soil and other environmental variables we used . . .”

P4 L22: What does the number 173 stand for? Is it the number of available covariates?

A: Yes, we might delete it to keep the sentence simpler. Number of predictors is mentioned elsewhere, it might not be necessary to highlight it also here.

P4 L30 – P5 L2: I had to read the sentences several times to understand them. Please rephrase.

A: We will rephrase those sentences: “For the construction of PTFs those samples were selected from the MARTHA dataset which have measured information on dependent and independent variables. We needed two kinds of predictions: (1) for topsoils where we could include organic matter content, calcium carbonate content and pH among the predictors and (2) for subsoils without the above soil chemical parameters, because those are not available for the 30-60 and 60-90 cm soil depths on the Balaton catchment. First we randomly selected 67% of the samples from those which have data on the dependent and all the independent variables available on the catchment

area to derive the PTFs. The rest 33% was used to compare the performance of the PTFs, this we called TEST_CHEM set. In the second step we needed a training (67% of data) and test set (33% of data) also for subsoil prediction for which we didn't have to apply the restriction on the soil chemical properties, therefore we could include more samples for the analysis. As test set we used the samples of the TEST_CHEM set and further added cases to reach the 33% of the complete data appropriate for subsoil predictions. Again the left 67% was used for the training. Number of samples used to train and test the PTFs was 8157 and 12039 for THS, 8051 and 11931 for FC, 8195 and 12036 for WP, with and without soil chemical properties respectively."

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.

A: Thank you for the suggestion. Information on how data was selected to train and test the PTFs will be moved under section 2.4.1. The heading will be rephrase, e.g.: Dataset to relate soil hydraulic properties and environmental information.

P5 L7: Please rephrase "most often used soil water retention values".

A: We will rephrase it. We mapped those soil water retention values which are usually differentiated for several applications.

P5 L7-8: Why did you map water content at -330 cm matric potential when field capacity is determined in Hungary at -300 cm?

A: Thank you for finding it, -300 is a mistyping error. It will be corrected in the text.

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.

A: We will rephrase the sentence, RF and GBM are two widely used MLA, which often achieve good prediction performance on datasets that are characterized by a large

number of predictors.

P5 L17: To calculate quantiles during the predictions? Quantiles of what? What is meant by "during the predictions"?

A: The text will be rephrased. "The advantage of these two algorithms is the possibility to estimate prediction intervals of the dependent variable as a function of the independent variables."

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.

A: We will summarize the principles of regression trees before describing the difference between RF and GMB.

P5 L19: . . .build ensembles of models. . .

A: Will be corrected.

P5 L19: . . . the difference between GBM and RF is the way. . .

A: Will be corrected.

P5 L26: mtry? Seems to be an argument of an R function?

A: Yes, we will clarify it in the text.

P6 L1: 50 independent variables out of how many? I assume that it is related to the number 173 in P4 L21. Right?

A: Yes for topsoils, in the case of subsoils it was 170, it will be clarified in the text.

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?

A: Thank you for highlighting it, the sentence was not properly phrased. Our aim was to reduce the number of predictors. We selected the 50-50 most important variables both in GBM and RF methods based on the five times repeated five-fold cross-validation, then concatenated the two sets of predictors. In this way less relevant predictors were excluded from the analysis. First we wanted to use the recursive feature elimination (Gregorutti et al., 2017) – with rfe function implemented in R caret package –, which would be a real optimization of input variable selection, but the RFE analysis couldn't be finished on the training set (173 variables of more than 5700 samples) due to lacking computation capacity. Then we found the possibility to at least reducing the number of predictors based on importance measure of the variables. Nussbaum et al. (2018) compared different covariate selection methods: a) based on variable importance calculated in RF model and b) stepwise recursive elimination of the least important variables. They found that both methods selected similar set of covariates. Their study was similar to our ones regarding the topic and dimension of data, therefore based on their results we reduced the number of predictors based on variable importance, which is practically the first step of the RFE analysis. In HUN-PTF method we considered the variable importance of both GBM and RF to rely on the results of two different methods. We concatenated the 50-50 most important variables, in this way depending on soil hydraulic parameter and soil depth 65-76 predictors stayed in the model. Text will be modified accordingly.

Gregorutti, B., Michel, B. and Saint-Pierre, P.: Correlation and variable importance in random forests, Stat. Comput., 27(3), 659–678, doi:10.1007/s11222-016-9646-1, 2017.

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.

A: Yes, we will add the term bootstrapping under 2.4.1. section.

P6 L18: . . .to the median and the 5% and 95% quantiles. . .

A: Thank you, we will modify it.

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.

A: Thank you for your observation. The essential difference between direct and indirect mapping is the approach of the inference. In direct mapping the target soil variable is directly interpolated over the domain of interest, whereas in indirect mapping not the target variable but its components / factors / covariates are interpolated first and then these interpolated surfaces are in use to compute and map the target variable. Pásztor et al. (2017) discussed this two approaches in detail.

P6 L26: . . . Table 2 summarizes the measured. . .

A: Thank you, we will modify it.

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".

A: Thank you for highlighting it. Yes, Table 1 shows all available environmental covariates, text will be corrected.

P7 L11: . . .with the method. . .

A: Thank you, we will modify it.

P7 L10: ". . .based on measured soil hydraulic properties calculated for. . .". How can the measured properties be calculated? Please rephrase the sentence.

A: It will be modified. We intend to say: Performance of soil hydraulic maps was eval-

uated based on observed soil hydraulic properties harmonized for 0-30, 30-60 and 60-90 cm depth with the method described in 2.4.2 section.

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?

A: We will add information on how number of variables was decreased under 2.4.1 section. Based on both GBM and RF analysis most important 50-50 variables was selected, after concatenating those we got 69-76 for topsoil predictions, for subsoils 65-77 stayed depending on the target variable in the case of HUN-PTF method.

P8 L9-16: This paragraph should be shifted to the Materials and Methods section.

A: The paragraph includes the result of tuning the model parameters before building the final model, therefore we thought to include it under the results of HUN-PTFs.

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.

A: Thanks to highlight it, the sentence will be revised. In case of our analysis multicollinearity is similar in the training set and mapped area therefore it had less influence on the performance of the maps. It is true that if the HUN-PTFs would be applied in a different region multicollinearities might influence the performance of the predictions. Dorman et al. (2013) found that prediction performance of random forest did not get worse due to high collinearity in the training dataset. The above mentioned RFE analysis would help to decrease multicollinearity (Gregorutti et al., 2017) but didn't run on our dataset due to high dimensionality. By eliminating around 100 predictors from the

entire 173 based on the importance measures we could partially decrease the multicollinearity and improve performance of the prediction. Optimizing predictor selection could be further elaborated, but this is beyond the aim of the presented paper.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D. and Lautenbach, S.: Collinearity: A review of methods to deal with it and a simulation study evaluating their performance, Ecography (Cop.)., 36(1), 027–046, doi:10.1111/j.1600-0587.2012.07348.x, 2013.

P8 L22: Please compare the values listed in the text and in Table 1 once again. I am not sure if they match.

A: Thank you, you are right, we will correct it. We will modify the R2 values as well, we will list those only for RF. In this way we keep the logic: highlighting results of the selected algorithm.

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.

A: Thanks for highlighting it. We used the environmental covariates as independent variables in HUN-PTF and the RF part of the RFK. For clarification "predictors" will also be used in text related to PTFs: under 2.1., 2.2, 3.1. Geostatisticians use the term "environmental covariate", therefore it might enhance the interpretability of the manuscript if also this term would be kept.

P8 L30: . . .than soil related variables. . .

A: Thank you, it will be corrected.

P9 L6-L26: Please explain in the Method section how relative importance is determined.

A: Explanation on it will be added.

P9 L30: mtry? See also my comment on P5 L26.

A: We will write it out: “number of randomly selected predictors at each split” for easier understanding.

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.

A: Thank you for the suggestion we will include the scatter plots after Fig. 2, which would also show the 90 % prediction intervals and describe it in the text. Please find it attached as P1., quality of figure will be increased: The scatter plot of the measured versus predicted water retention values with 90% prediction interval on test data sets based on random forest method. THS: saturated water content, FC: water content at field capacity, WP: water content at wilting point.

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.

A: You are right in that sense kriged values computed with a pure nugget model do not give any new “information” to the RF predictions. However, kriged values with a nugget variogram add zero values to the RF predictions rather than random numbers. Thus kriging with a nugget model do not blur the predictions. We would not like to leave this exercise because it is an algorithmical decision (considering the stochastic part of the spatial variation of the given soil property) rather than a subjective decision, even if we get the same result.

P10 L10: the correlation is based on only three pairs of values. Please use a weaker formulation.

A: Thank you, we will rephrase the sentence.

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.

A: We will show the Fig. 5 and 6 maps also for THS and FC, please find those attached (P2, P3, P4). Sorry for confusion, the formulation of the aim of the paper will be identical in the entire manuscript. Calculating the confidence intervals for the RFK method is beyond the scope of this study, although it would be interesting to analyze the difference between uncertainty maps calculated with the different methods in the future, similarly as it was done by Szatmári and Pásztor (2018) for soil organic carbon stock in Hungary. According to it, quantile regression forest (Meinshausen, 2006) based uncertainty quantification outperforms most of the prediction techniques used in digital soil mapping. Furthermore, they have pointed out that bootstrapping based uncertainty quantification for RFK is quite time consuming, as well as it requires massive storage and computing capacity. The ranger package - with which we derived the HUN-PTFs - includes implementation of quantile regression forest (Meinshausen, 2006) for the calculations of the prediction intervals.

Meinshausen, N.: Quantile Regression Forests, J. Mach. Learn. Res., 7, 983–999, 2006.

P11 L12: . . .we have not differentiated uncertainty of. . .

A: Thank you, it will be corrected.

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.

A: We were not aware of the correct formulation of the conclusions and wrongly included discussion in that section. The text will be completely moved under Results and discussion section and we will add the real conclusions there. The take home message is the following: Based on the results in case of six out of nine soil hydraulic maps there is no significant difference in performance between values derived with pedotransfer function and geostatistical method on the Balaton catchment. The benefit of maps derived with random forest and kriging is that locally extreme values can be better characterized. In the case of pedotransfer function based mapping it is advantageous that calculation of uncertainty is much less computation intensive than it is with geostatistical methods, although it would be interesting in the future to analyze the difference between uncertainty maps calculated with the different methods specifically for soil hydraulic properties.

P12 L30-32: A discussion of methods or suggestions of alternative methods should be done in the discussion section.

A: It will be moved under Results and 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.

A: Both text and Table 7 will be moved under Results and discussion 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)).

A: Text will be moved under Results and discussion section.

Thank you again the comments and suggestions. We hope to adequately address the issues identified and look forward to any other feedback the referee may have.

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

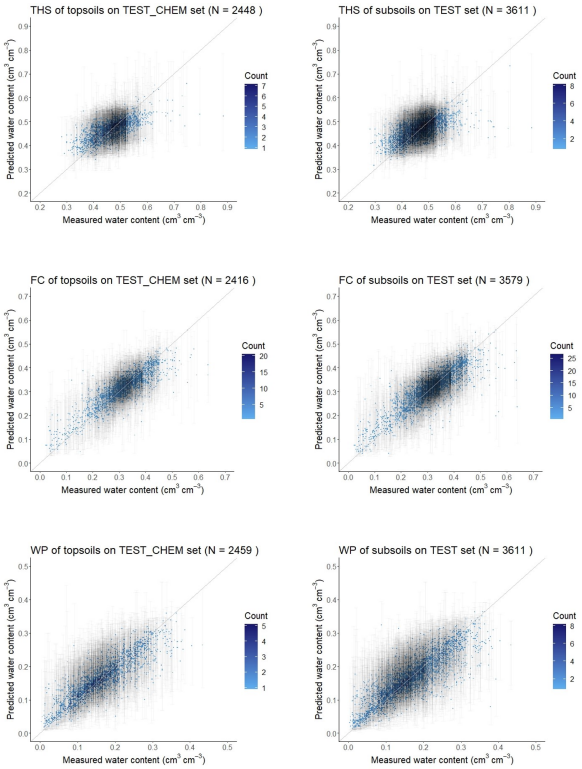

Fig. 1. P1

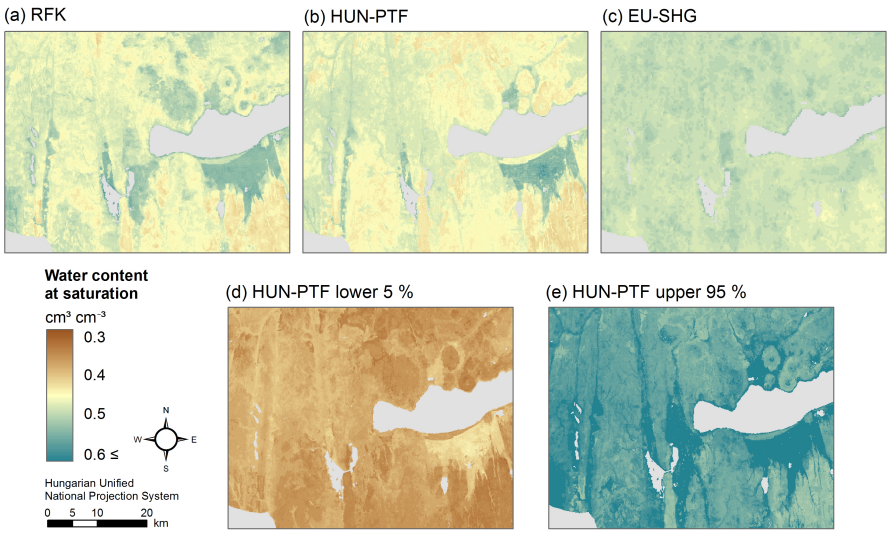

Fig. 2. P2

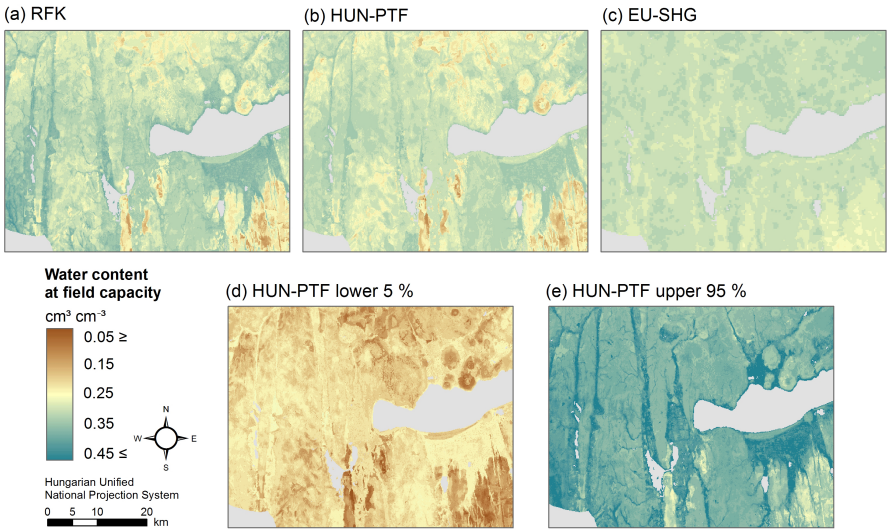

Fig. 3. P3

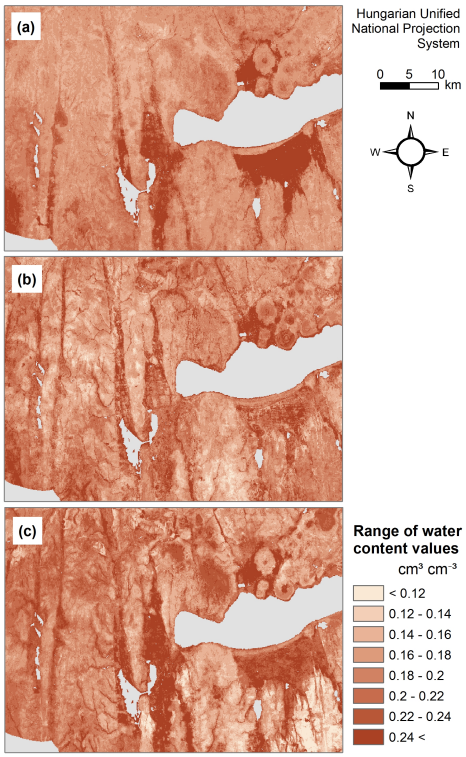

Fig. 4. P4