Revison of manuscript on “Mapping soil hydraulic properties using random forest based pedotransfer functions and geostatistics”

Content of the document

I. POINT BY POINT AUTHORS' RESPONSE TO THE REVIEWERS

1. RESPONSE TO THE EDITOR

Dear Dr. Ehret,

Thank you for the positive evaluation of our replies and highlighting those three important points which will significantly improve the clarity of the manuscript. Please find here our answers following the raised points with page and line numbering of the revised manuscript:

Q1: Language: Although there will be copy-editing done by HESS after the manuscript has been accepted, please have a native speaker read your manuscript to check for spelling, grammar and readability.

A1: The whole manuscript has been checked by an English language expert.

Q2: Conclusions: As referee #1 mentioned, a conclusions section should be short and concise. It should pick up the main questions/hypotheses of the manuscript, provide the related findings of the paper and point to deficiencies and required future work.

A2: We were not aware of the correct formulation of the conclusions and wrongly included discussion in that section. The text has been completely moved under Results and discussion section – to P14 L12 - P15 L16 – and we added the real conclusions there – P15 L23-28.

Q3: As mentioned by referee #1 (P6L20-P7L8), please make clear the difference between your two main approaches, and why you call one 'direct' and the other 'indirect'. If this is made clear from the very beginning of the paper, it will be much easier to understand. I recommend adding an illustrative flowchart explaining the sequence of steps of each method.

A3: Thank you for highlighting that the difference between the methods was not completely clear and giving the idea of the flowchart. We have added a paragraph under 2.4 section – P6 L5-13 – , in which we summarize the main steps of the two methods and added the flowchart on it as Fig. 2.

We hope we could properly address all the questions in the revised version of the article.

With regards,
Authors

2. RESPONSE TO REFEREE #1

Dear Referee #1,

Thank you for the detailed review and suggestions which help us to improve our manuscript. We have tried to address all the comments in the revised version of the article. Below we would like to answer the questions and recommendations, following the general and special comments. **Page and line numbering refer to that of the revised manuscript with track changes.**

General comments:

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

A1: 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 and revised description on it on P7 L22-28.

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.

A2: Thank you for calling our attention to reformat the conclusions. Information and Table 7 have been moved under results and discussion to P14 L12-P15 L16 and conclusions have been rewritten on P15 L23-28. Please find more information on it under answers for specific comments.

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

A3: We are sorry that the download link of the maps (https://www.mta-taki.hu/en/kh124765/maps) did not work, something happened with the access authorization after submitting the manuscript, now the problem has been fixed. We also added a link to download the HUN-PTFs: https://www.mta-taki.hu/en/kh124765/hun_ptfs on P16 L29-P17 L2.

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 has been added in the abstract as well: P1 L18-19.

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 have provided some papers, which used several MLAs (e.g. neural networks, cubist, gradient boosting) for mapping soils and short summary of them: P3 L4-11.

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

A: Thank you for highlighting it, it has been clarified: P3 L22

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 has been clarified accordingly: P3 L23-24.

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. We have added this information in the text on P3 L33-P4 L7 and 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 have rephrased the objective mentioned in the text to be identical. We 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.

Please find changes in text on P4 L9-12.

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 decreased the number of abbreviations, MLA, OM, $CaCO_3$ and EU-SHG are not used in the text and unnecessary ST and PSD abbreviations have been removed too (eg.: P2 L12, L24; P5 L1; P4 L27-30, etc.). The THS, FC, WP and MARTHA abbreviations have been deleted from the abstract: P1 L18-9, 23, 27.

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".

A: Thank you for highlighting it, the sentence has been rephrased on P5 L7-8.

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

A: Yes, we have deleted it to keep the sentence simpler on P5 L8. 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 have rephrased those sentences on P6 L18-30

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 has been moved under section 2.4.1. (P6 L18-30) The heading has been rephrased on P5 L6: Dataset to relate soil hydraulic properties and environmental information.

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

A: We have rephrased it: "We mapped soil water content at 0, -330 and -15,000 cm matric potential values, THS, FC and WP respectively, because these soil hydraulic properties are often required for various purposes".

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 was a mistyping error. It has been corrected in the text on P5 L28.

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 have rephrased the sentence on P7 L2-4. 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 has been rephrased on P7 L4-5.

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 have summarized the principles of regression trees before describing the difference between RF and GMB on P7 L7-15

P5 L19: …build ensembles of models…

A: It has been corrected on P7 L7.

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

A: It has been corrected on P7 L8-9.

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

A: Yes, we have clarified it in the text on P7 L18.

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 has been clarified in the text on P7 L26-27

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 has been modified accordingly on P7 L22-28.

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 have added the term bootstrapping under 2.4.1. section on P7 L10 and P8 L3.

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

A: Thank you, we have modified it on P8 L12-13.

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. We have added a paragraph under 2.4 section – P6 L5-13 – , in which we summarize the main steps of the two methods and added a flowchart as Fig. 2, which highlights the difference between the direct and the indirect method. 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 have modified it on P8 L21-22

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 has been corrected on P8 L27.

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

A: Thank you, we have modified it on P9 L5.

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

A: It has been modified on P9 L4-5.

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 have added information on how number of variables was decreased under 2.4.1 section on P7 L27-28.

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 the tuning of the model parameters before building the final model, therefore we thought to include it under the results of HUN-PTFs on P10 L11-18.

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: Thank you to highlight it, the sentence has been revised and complemented on P10 L2-9.

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 have corrected it on P10 L23-27. We have modified the R^2 values as well, we listed those only for RF. In this way we keep the logic: highlighting results of the selected algorithms.

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 we introduced the term "predictors" in text related to PTFs: under 2.1. (P5 L1), 2.4. (P6 L20; P7 L11, 15, 23, 24, 26, 28; P8 L8, 28), 3.1. (P10 L2, 32), 3.2 (P12 L7). 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 has been corrected on P11 L2.

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

A: Explanation on it has been added, on P7 L23-25.

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

A: We have written it out on P12 L7: "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 have added the scatter plots in Fig 4, which shows the 90 % prediction intervals and refer it in the text on P10 L27-30.

Figure 4. 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, TEST_CHEM set: test dataset in which chemical soil properties are available for the predictions, TEST set: test dataset, in which chemical soil properties are not necessarily available for the predictions.

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 have rephrased the sentence on P12 L18-19.

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 have added maps also for THS and FC, please find those as Fig 7, 8, 10. Text has been modified accordingly on P13 L10-12, 19.

Sorry for confusion, the formulation of the aim of the paper has been cleared 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.

Information related to bootstrapping based uncertainty quantification for RFK has been added on P15 L8-12.

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 has been corrected on P13 L23-24.

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 has been completely moved under Results and discussion section to P14 L12-P15 L16, and we have added the real conclusions on P15 L23-28. The take home message is the following:

Based on results of six out of nine soil hydraulic maps there is no significant difference in performance between values derived using pedotransfer function and geostatistical method on the Balaton catchment area. The benefit of maps computed with random forest and kriging is that locally extreme values can be characterized better. 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 analyse 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 has been moved under Results and discussion section to P15 L4-6.

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 have been moved under Results and discussion section to P15 L13-15.

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 has been moved under Results and discussion section to P15 L15-16.

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

With regards,
Authors

3. RESPONSE TO REFEREE #2

Dear Referee #2,

Thank you for the review and suggestions for clarification, which helps to improve the quality of the manuscript. We hope that we could address all the raised questions and comments in the revised version of the manuscript. Please find our answers for the questions and recommendations, following the general and specific comments. **Page and line numbering refer to that of the revised manuscript with track changes**.

General comments

Q1: This is an interesting manuscript investigating an important topic. The manuscript is well structured, but it is difficult to follow in places. Especially the methods need further clarification and details.

A1: Thank you for the review and suggestion for clarification. The entire manuscript has been checked by a specialized language expert to improve intelligibility of the text. The followings have been modified/included related to description of the methods:

- information and flowchart (Fig.2) has been added to clarify difference between direct and indirect methods on P6 L5-13,
- the number of abbreviations has been decreased, MLA, OM, $CaCO_3$, EU-SHG, ST and PSD have been removed from the text (eg.: P2 L12, L24; P5 L1; P4 L27-30, etc.)
- data partition to train and test the pedotransfer functions has been rephrased on P6 L18-30,
- information on how data was selected to train and test the PTFs has been moved under section 2.4.1. and rephrased (P6 L18-30),
- the heading has been rephrased on P5 L6: Dataset to relate soil hydraulic properties and environmental information, and only information about the soil hydraulic dataset has been kept there (P5 L7-13),
- the principles of regression trees have been summarized before describing the difference between RF and GMB on P7 L7-15,
- meaning of mtry has been clarified on P7 L18 and P12 L7,
- information on how number of variables was decreased has been rephrased on P7 L22-28 to increase clarity,
- the sentence about multicollinarity has been revised and complemented on P10 L2-9.

Specific comments

Q1: Lines 11-13: very long sentence

A1: Thank you for highlighting it. We have rephrased the first part of the abstract on P1 L10-15.

Q2: Lines 13-16: not clear

A2: We have structured the sentence on P1 L15-16 and rephrased the sentences on P1 L15-18.

Q3: Abstract: in general: a lot of acronyms for an abstract. Not clear and difficult to understand what direct and indirect are.

A3: We have eliminated the following abbreviations from the abstract: THS, FC, WP, MARTHA. In the direct method we used the geostatistical approach to spatially inference measured soil hydraulic data collected in profiles of the catchment through modelling its relationship with environmental covariates. In indirect mapping PTFs were derived first to describe relationships between soil hydraulic properties and easily available soil and other environmental parameters. The PTF predictions were then spatially implemented on the environmental covariates clipped for the catchment area of Lake Balaton. We have added some small edits in the abstract to clarify it:

- on P1 L15-16: "… soil hydraulic properties derived from (i) indirect (using PTFs) and (ii) direct (geostatistical) mapping methods" and

- on P1 L22-23: “As a direct, thus geostatistical method random forest combined with kriging (RFK) was applied …”

Q4: Introduction: I think a language revision is needed. The first paragraph for example is difficult to understand.

A4: Thank you for the suggestion. Along the suggestions of the two reviewers we made a lot of corrections for the improvement of the manuscript. The language of the manuscript has been edited by a language expert.

Q5: Lines 19-22: this is not always true. Please discuss.

A5: Thank you for highlighting it. The sentence has been rephrased on P2 L21-23.

Q6: Section 2.2: the description of the soil hydraulic dataset is not clear. The split between test and training in particular.

A6: Thank you for highlighting it. We have rephrased the description on splitting the dataset into train and test on P6 L18-30. We have moved that paragraph under section 2.4.1 to clarify that this splitting was performed for the HUN-PTF approach.
Under section 2.2 (P5 L6-13) we have only kept information about the soil hydraulic dataset, and added some clarification about what kind of data is included in the dataset on P5 L11-13, and rephrased the title of the section (P5 L6).

Q7: The results are well presented, but there are minor (and less minor) problems with typos and structure of the sentences.

A7: Thank you. We have corrected the typos related to RMSE and R2 values on P10 L23-26. Further typos and structure of the sentences has been corrected by a language expert.

Thank you for the review. We hope that we could clarify text related to the methods and improve the intelligibility of the entire manuscript. We are looking forward the feedback about the revised manuscript.

With regards,
Authors

II. LIST OF ALL RELEVANT CHANGES MADE IN THE MANUSCRIPT

Page and line numbering refer to that of the revised manuscript with track changes.

The following changes have been made in the manuscript:

- the English has been edited in the entire manuscript;
- the following abbreviations have been removed from the text of the manuscript: MLA, OM, $CaCO_3$, EU-SHG, ST and PSD;
- for clarification we introduced the term “predictors” in text related to PTFs: under 2.1. (P5 L1), 2.4. (P6 L20; P7 L11, 15, 23, 24, 26, 28; P8 L8, 28), 3.1. (P10 L2, 32), 3.2 (P12 L7)
- P1 L4: the first author will use her maiden name: Brigitta Szabó;
- abstract has been rephrased to clarify the text and decrease abbreviations: P1 L10-19, L21, L23, L27;
- introduction:
    - P2 L21-23: text has been rephrased and a sentence has been added for clarification;

- P3 L1-2: tested machine learning algorithms are listed,
- P3 L4-11: short summary of papers has been added, which used several machine learning algorithms (e.g. neural networks, cubist, gradient boosting) for mapping soils;
- P3 L22-24: two sentence have been clarified;
- P3 L33-P4 L7: published numbers on the magnitude of uncertainties related to the prediction of soil water retention have been added;
- P4 L9-12: aim of the study has been clarified;

– materials and methods:
- P5 L6: title of section has been modified to better describe its content;
- P5 L7-8: sentence has been clarified;
- P5 L11-13: information about the properties included in the MARTHA dataset has been added;
- P5 L14-24: text has been moved under 2.4.1 to P6 L15-30 to clarify that the dataset was divided in a certain way only in the case of the HUN-PTF method, the text has been rephrased to increase intelligibility;
- P6 L4: title has been modified;
- P6 L5-13: a summary about the main steps of the two methods and a flowchart on it as Fig. 2 has been added;
- P7 L1: sentence has been rephrased;
- P7 L4-5: sentence on computing uncertainty has been clarified;
- P7 L7-15: summary on the principles of regression trees has been added;
- P7 L10, P8 L3: the term bootstrap sample is used for clarification;
- P7 L18: meaning of argument mtry has been clarified;
- P7 L21-28: description about the selection of predictors has been rephrased;
- P8 L12-14: terms have been corrected and sentence has been rephrased;
- P8 L27: clarification has been added;
- P9 L4-5: sentence has been clarified;
- P9 L19: reference of R software has been added through Mendeley's Word Plug In;

– results and discussion:
- P9 L28-P10 L10: text related to variable selection and multicollinearity has been completed;
- P10 L24-26: typos have been removed;
- P10 L27-30: scatterplot – as Fig. 4 – and its description has been added;
- P12 L7: mtry has been replaced with its description;
- P12 L18-19: the sentence has been rephrased;
- P13 L12-14: maps of THS and FC has been added on separate figures: Fig. 7, 8, and added to Fig. 10;
- P14 L11-P15 L16: a new section has been introduced, text and Table 7 from previous version of the conclusions (P15 L29-P6 L27) have been moved here;
- P15 L8-12: information related to bootstrapping based uncertainty quantification for RFK has been added;

– conclusions
- P15 L23-28: conclusions has been added, text related to the discussion (P15 L19-23 and P15 L29-P16 L27) has been removed as mentioned before;

– data availability:
- P16 L29-P17 L2: link to download HUN-PTFs has been added;

– references: new reference has been added with Mendeley's Word Plug In – therefore not highlighted with track changes – on
- P18 L3-4,
- P18 L31-34,
- P19 L13-14,
- P19 L17-19,

    - P20 L16-18,
    - P20 L31-33,
    - P21 L11-13,
    - P21 L34-P22 L1
- tables:
    - P26 L3-4: information has been added on test sets;
    - P26 L5: information on soil type has been clarified;
- figures: numbering of figures has been revised due to adding four more figures,
    - P33: flowchart (Fig. 2) has been added, which describes direct and indirect method;
    - P35: scatterplot (Fig. 4) has been added, which shows measured vs predicted values with 90% prediction intervals;
    - P37 L4-5: sentence has been clarified;
    - P38: map of THS of 0-30 cm soil depth has been added (Fig. 7);
    - P39: map of FC of 0-30 cm soil depth has been added (Fig. 8);
    - P40-41: map of WP of 0-30 cm soil depth has been replaced (Fig. 9) and its caption has been clarified;
    - P42-43: figure has been replaced by map also including THS and FC (Fig. 10).

III. MARKED-UP MANUSCRIPT VERSION

Please find revised marked-up manuscript on the following pages.

Mapping soil hydraulic properties using random forest based pedotransfer functions and geostatistics

Brigitta Szabó~~Tóth~~[1,2], Gábor Szatmári[1], Katalin Takács[1], Annamária Laborczi[1], András Makó[1], Kálmán Rajkai[1], László Pásztor[1]

[1]Institute for Soil Sciences and Agricultural Chemistry, Centre for Agricultural Research, Hungarian Academy of Sciences, Herman Ottó út 15, 1022 Budapest, Hungary
[2]Georgikon Faculty, University of Pannonia, Deák Ferenc u. 16, 8360 Keszthely, Hungary

Correspondence to: Gábor Szatmári (szatmari.gabor@agrar.mta.hu)

Abstract. Spatial 3D information on soil hydraulic properties for areas larger than plot scale are usually derived ~~with~~ using indirect methods such as pedotransfer functions (PTFs) due to ~~lacking~~ the lack of measured information on them~~those~~. ~~Soil hydraulic properties are calculated with applying pedotransfer functions (PTFs) —~~ PTFs ~~which~~ describe the relationship between the desired soil hydraulic parameter and easily available soil properties ~~determined~~ based on a soil hydraulic ~~point~~ reference dataset. Soil hydraulic properties of a catchment or region can be calculated by applying PTFs~~—~~ on available soil maps. Our aim was to analyse ~~difference in~~the performance and spatial ~~patterns~~ distribution ~~between~~ of soil hydraulic ~~maps~~ properties derived ~~with~~from (i) indirect (using PTFs) and (ii) direct (geostatistical) mapping methods. ~~We performed the~~A study was performed on the Balaton catchment area in Hungary, where density of measured soil hydraulic data fulfils the requirements of geostatistical methods. Maps of saturated water content (~~THS~~0 cm matric potential), field capacity (~~FC~~-330 cm matric potential) and wilting point (~~WP~~-15000 cm matric potential) for 0-30, 30-60 and 60-90 cm soil depth were prepared. PTFs were derived ~~with~~ using the random forest method on the whole Hungarian soil hydraulic dataset, which includes ~~(MARTHA:~~ soil chemical, physical, taxonomical and hydraulic ~~information~~ properties of some 12,000 samples~~)~~ complemented with information on topography, climate, parent material, vegetation and land use. As a direct, thus geostatistical method random forest combined with kriging (RFK) was applied ~~on~~ to 359 ~~MARTHA~~ soil profiles located in the Balaton catchment area. There ~~was~~ were no significant differences between the direct and indirect methods in ~~case of~~ six out of nine maps having root mean squared error values between 0.052 and 0.074 $cm^3\ cm^{-3}$, which is in accordance with the internationally accepted performance of hydraulic PTFs. The PTFs based mapping method performed significantly better than the RFK for the ~~THS~~ saturated water content at 30-60 and 60-90 cm soil depth, in the case of ~~WP~~ wilting point the RFK outperformed the PTFs at 60-90 cm depth. Differences between the PTF based and RFK mapped values are less than 0.025 $cm^3\ cm^{-3}$ for 65-86 % of the catchment. In RFK, uncertainty of input environmental covariate layers is less influential on the mapped values which is preferable. In the PTFs based method the uncertainty of mapping ~~of the~~ soil hydraulic properties is

less computational intensive. Detailed comparisons of ~~the~~ maps derived ~~by~~ from the PTF based method and the RFK ~~is~~ are presented in ~~the~~ this paper.

1 Introduction

Providing information on soil hydraulic properties is desired for many environmental modelling studies (Van Looy et al., 2017). Most often, measured information on soil water retention or hydraulic conductivity is not ~~even~~ available for ~~small water catchments~~environmental modelling ~~n~~either at regional or continental scale. Analyses on the prediction of soil hydraulic properties ~~has been~~were started extensively in the 1980s (Ahuja et al., 1985; Pachepsky et al., 1982; Rawls and Brakensiek, 1982; Saxton et al., 1986; Vereecken et al., 1989) and are continuously updated to increase the performance of predictions (pedotransfer functions - PTFs) when newer statistical methods and/or new data become available. Latest works ~~on it~~include among others ~~are~~ McNeill et al. (2018); Román Dobarco et al. (2019); Zhang and Schaap (2017).

Tree based machine learning algorithms ~~(MLA)~~ have been found to be efficient tools in general for predicting purposes (Caruana et al., 2008; Caruana and Niculescu-Mizil, 2006; Olson et al., 2017), especially gradient tree boosting and random forest. These methods are used to derive ensembles of trees, providing predictions of several individual trees ~~built~~ with built in randomization. Tree type algorithms provide mean values of groups that can be statistically differentiated, called terminal nodes (Breiman, 2001). Due to this way of providing estimations, these methods do not derive any extraordinary values, therefore predictions will ~~be~~ always be reasonable if training data is appropriately cleaned. For the same reason it decreases variability as well, extreme values are smoothed out (Hengl et al., 2018b).

Ensemble predictions can be derived not only ~~by~~ from a single ~~MLA~~method, which consist of several models through bagging or boosting of e.g. decision tree, or support vector machine, or neural network algorithms, but can consist of different models and is derived from the average of all. It has been shown~~,~~ that often but not always, the more models are combined for the prediction the more accurate the results are ~~is~~ (Baker and Ellison, 2008; Cichota et al., 2013; Nussbaum et al., 2018; Wu et al., 2018). Although the significancy of improvement is often not tested. Hengl et al. (2017) also used merged ensemble predictions by calculating the weighted average of two machine learning algorithm~~MLA~~s to decrease influence of model overfitting. Although from the application point of view it is important to avoid increasing the complexity and size of the prediction model if there is no significant improvement in performance. Accuracy, interpretability and computation power required to use the prediction algorithm have to be optimized at the same time for allowing widespread use of derived models.

Tree type ensemble algorithms were found to be successful ~~for~~ in harmonizing different soil texture classification systems (Cisty et al., 2015), prediction of soil bulk density (Chen et al., 2018; Dharumarajan et al., 2017; Ramcharan et al., 2017; Sequeira et al., 2014; Souza et al., 2016), but ~~were~~ have not been ~~yet~~ intensively applied yet to derive input parameters for hydrological modelling (Koestel and Jorda, 2014; Tóth et al., 2014).

Hengl et al. (2018a) tested several machine learning algorithm~~MLA~~s (i.e. neural networks, random forest, gradient boosting, K-nearest neighbourhood and cubist) to map potential natural vegetation. From those random forest performed the best. Nussbaum et al. (2018) analysed different methods to map several soil properties for three study sites in Switzerland. They also found that the random forest method performed the best when a single model ~~is~~ was used. Adhikari et al. (2014) used the cubist method 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 ~~MLA~~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. Results of Rudiyanto et al. (2018) also showed that among several tested ~~MLAs~~ methods tree-based models performed the best. Hengl et al. (2018b) reviewed machine learning algorithm~~MLA~~s and geostatistical methods for soil mapping and found that the random forest method combined with the calculation of geographical proximity effects is a powerful method similarly to universal kriging.

Soil hydraulic maps are mostly derived ~~in~~ by two ways i) by applying pedotransfer functions (PTFs) on available soil and/or environmental maps, called ~~as~~ an indirect mapping method, ii) with direct spatial inference of observation point data (Bouma, 1989), which is considered ~~as~~ to be a direct procedure. Point data can be measured or predicted by PTFs. Several studies analysed the efficiency of geostatistical methods to map water retention at specific matric potential (Farkas et al., 2008) and saturated hydraulic conductivity (Motaghian and Mohammadi, 2011; Xu et al., 2017). Ferrer Julià et al. (2004) mapped soil hydraulic conductivity for the Spanish area of the Iberian Peninsula at 1 km resolution with both methods (i) and (ii). They found that the map derived by kriging interpolation performed the best. Farkas et al. (2008) mapped water content at field capacity and wilting point with geostatsitical methods for an area of 1483 ha. They optimized ~~number of measurements~~sampling density needed to derive 10 m resolution soil hydraulic maps for their study site.
In most ~~of the~~ cases there is no available point data for applying geostatistical methods, therefore in several studies soil hydraulic maps were generated with a PTF applied on easily available spatial soil data (Chaney et al., 2016; Dai et al., 2013; Marthews et al., 2014; Montzka et al., 2017; Tóth et al., 2017; Wu et al., 2018).
Further to the spatial variability of soil hydraulic properties, information on the prediction uncertainty is important for modelling tasks. In this way extreme conditions might be better described. A ~~P~~possible calculation of this kind of uncertainty was provided by Montzka et al. (2017). They calculated sub-grid variability of the coupled Mualem-van Genuchten model parameters for a coarse 0.25° grid based on fitting water retention and hydraulic conductivity model for each grid cell of the 1 km resolution SoilGrids. Román Dobarco et al. (2019) and McNeill et al. (2018) also provided information on the uncertainty of the prediction of soil hydraulic properties. Root mean squared error (RMSE) of published PTFs predicting soil water retention is usually between 0.02 and 0.07 cm^3 cm^{-3} depending on the predicted soil hydraulic property and available input

information, e.g. in (Nguyen et al., (2017), (Zhang and Schaap, (2017) or (Román Dobarco et al., (2019) to mention some of the latest results. When PTFs are used for mapping, the uncertainty of the input soil layers will further increase the uncertainty of the calculated soil hydraulic properties, e.g. in point based validation RMSE was 0.073 cm^3 cm^{-3} for water content at field capacity mapped for China in (Wu et al., (2018); (Leenaars et al., (2018) found that mean RMSE for water content at saturation, field capacity and wilting point together was 0.102 cm^3 cm^{-3} for African soils; in EU-SoilHydroGrids (Tóth et al., 2017) RMSE was 0.095, 0.096, 0.084 cm^3 cm^{-3} for water content at saturation, field capacity and wilting point respectively for European soils.

Our aim was twofold, 1) to analyse how different mapping methods could be applied to derive maps of soil hydraulic properties, such as water content at saturation (THS), field capacity (FC) and wilting point (WP) and 2) provide a non-computation intensive method to map uncertainty of calculated soil hydraulic parameters on the Balaton catchment area in Hungary. Soil hydraulic maps were derived by i) an indirect method: applying local hydraulic PTFs on the available soil and other environmental spatial information of the catchment and ii) geostatistical – direct – method using available soil profile data and environmental covariates of the catchment. Performance of derived soil hydraulic maps was compared with to that of the 3D European soil hydraulic maps (EU-SoilHydroGrids v1.0) (Tóth et al., 2017).

2 Materials and methods

2.1 Study site

We selected the catchment area of Lake Balaton (Fig. 1) to study mapping of soil hydraulic properties, because it is an important area in Hungary from the point of modelling hydrological, ecological, meteorological processes or planning land use and management. The size of the catchment is 5775 km^2. The mean depth of the lake is 3.5 m therefore water quality and quantity of the lake is sensible forsensitive to environmental changes. It has a warm temperate climate with 9-12°C mean annual temperature and 560-770 mm mean annual precipitation, lower temperature and higher rainfall values tend to be towards the western and hilly partselevated areas. Elevation is between 100 and 500 m on the northern part and 100 and 300 m in other areas of the catchment. Main soil types are Luvisols (53%), Cambisols (18%), Gleysols (10%), Histosols (5%) further to those Stagnosols, Arenosols, Regosols, Leptosols and Chernozems also occur (IUSS Working Group WRB, 2014).

For the catchment spatial information on soil type (ST), clay, silt and sand content (PSD), organic matter content (OM), calcium carbonate content ($CaCO_3$) and pH in water (pH) at 100 m resolution was provided by the DOSoReMI.hu (Digital, Optimized Soil Related Maps and Information; (Pásztor et al., 2018b)) framework (Table 1). Actually,As soil chemical properties – OMorganic matter content, calcium carbonate content$CaCO_3$ and pH – were only available only for the 0-30 cm depth, therefore those could only be considered only for the topsoil predictions. Information on topography, meteorology, geology

and vegetation listed in Table 1 was used as predictors and environmental covariates for the elaboration of PTFs ~~as well as for~~and direct mapping accordingly.

Topographical parameters were calculated with SAGA GIS tools (Conrad et al., 2015) based on the digital elevation model. For the mapping of soil hydraulic properties all covariates were harmonized, projected to the Hungarian Uniform National Projection system, rasterized if necessary and resampled to 100 m resolution.

2.2 ~~Soil hydraulic d~~Dataset to relate soil hydraulic properties and environmental information

For the ~~analysis of the relationship between~~prediction of soil hydraulic properties ~~and~~ based on soil and other environmental ~~covariates~~variables ~~(173) we used~~ the Hungarian Detailed Soil Hydrophysical Database (Makó et al., 2010) was used, extended with topographical, meteorological, geological information and remotely sensed vegetation properties (Table 1), called MARTHA ver 3.0 (acronym of the Hungarian name of the dataset). MARTHA consists of 15142 soil horizons' data belonging to 3970 soil profiles. The samples in it have measured information on basic soil properties – e.g. soil depth, organic matter content, clay, silt and sand content, calcium carbonate content, pH, etc. – and also on soil hydraulic properties such as soil water retention at different matric potential values.

~~We derived PTF for THS, FC and WP using soil depth, soil properties, environmental covariates listed in Table 1 as independent variables. OM, $CaCO_3$ and pH could be considered only for the topsoil (0-30 cm) predictions.~~

~~For the construction of PTFs those samples were selected from MARTHA which have measured information on dependent and independent variables. The dataset was randomly split into training sets to derive the PTFs and test sets to compare the performance of the PTFs. Two training and test sets were selected sequentially. One applicable for both the top- and subsoil predictions, the other only for topsoil estimations. First we randomly choose 33% of data applicable to test topsoil PTFs (TEST_CHEM). Then the other set (TEST), which included all the TEST_CHEM samples and further as many samples as were needed to reach the random 33% of all the data without chemical properties. In this way ratio of training and test sets were 67 and 33% respectively for each soil hydraulic predictions. Number of samples used to derive 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.~~

2.3 Mapped soil hydraulic properties

We mapped ~~the most often used soil water retention values,~~soil water content at 0, -330 and -15,000 cm matric potential values, THS, FC and WP respectively, because these soil hydraulic properties are often required for various purposes. Definition of FC varies across different countries. In Hungary FC is determined at ~~-300~~ 330 cm matric potential, therefore water content at -100 or -200 cm was not analysed in the presented work.

The information on soil properties were available for 0-30, 30-60 and 60-90 cm soil depths and this determined the vertical resolution of the soil hydraulic maps. As PTFs include depth as independent variable, they are applicable for any soil depth intervals.

2.4 Methods for soil hydraulic properties mapping

Soil hydraulic properties were mapped both with direct and indirect methods for the catchment of Lake Balaton. 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, and/or covariates are interpolated first and then these interpolated surfaces are used to compute and map the target variable. In the direct method we used the geostatistical approach to spatially inference measured soil hydraulic data collected in profiles of the catchment through modelling its relationship with environmental covariates. In indirect mapping PTFs were derived first to describe relationships between soil hydraulic properties and easily available soil and other environmental parameters. In this approach the full national MARTHA database provided soil reference data, and nationwide, spatially exhaustive environmental auxiliary information was used. The PTF predictions were then spatially implemented on the environmental covariates clipped for the catchment area of Lake Balaton (Fig. 2).

2.4.1 Pedotransfer function based indirect mapping (HUN-PTF)

We derived PTFs for THS, FC and WP using soil depth, soil properties and other environmental covariates listed in Table 1 as independent variables. Organic matter content, calcium carbonate content and pH could be considered only for the topsoil (0-30 cm) predictions, because those are not available for the subsoils on the Balaton catchment area.

For the construction of PTFs those samples were selected from the MARTHA dataset which had measured values of soil horizons or layers considered as 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 had data on the dependent and all the independent variables available on the catchment area to derive the PTFs. The remaining 33% was used to compare the performance of the PTFs, this we called TEST_CHEM set. In the second step we needed a training set (67% of data) and a test set (33% of data) also for subsoil prediction for which we did not have to apply the restriction on the soil chemical properties, therefore we could include more samples for the analysis. As a 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 remaining 67% was used for training.

The number of samples used to train and test the PTFs was 8,157 and 12039 for THS, 8,051 and 11,931 for FC, 8,195 and 12,036 for WP, with and without soil chemical properties respectively.

We analysed prediction performance of the two widely used~~most efficient~~ machine learning algorithms~~MLAs~~, ~~namely~~ random forest (RF) of R package 'ranger' (Wright, Wager, & Probst, 2018) and generalized boosted regression model (GBM) of 'gbm' (Ridgeway, 2017) for the prediction of THS, FC and WP. The advantage of these two algorithms is that the~~the possibility to~~ prediction intervals of the dependent variable are computed as a function of the independent variables~~calculate quantiles during the prediction, in this way prediction uncertainty can be provided based on parameter input combination~~.

Both algorithms build ensembles of models from regression trees. In regression trees data is recursively partitioned to increase homogeneity in the subsets, in this way residual sum of squares are minimized (Breiman et al., 1984). The difference between GBM and RF is the way ~~of building~~ the forest is built from the individual trees. RF relies on averaging the result of the trees in the ensemble. The trees are grown on n_{tree} bootstrap samples of the training data independently from each other (Breiman, 2001), therefore it is a bagging type ensemble. At each split of the trees only a small set of predictors is selected randomly to analyse which variable at which split point is the best for the partition, i.e. minimize the sum of squares. In GBM the ensemble model is grown sequentially, at each iteration step the next model is built with respect to the error of the ensemble learnt so far (Friedman, 2001; Natekin and Knoll, 2013), which is characteristic for the boosting type ensemble, already included in its name (Dietterich, 2000). In each split all possible predictors are considered.

Optimization of parameter set in RF and GBM model was performed with the train function of R package 'caret' (Kuhn et al., 2018). Five times repeated five-fold cross-validation was used to evaluate performance of different parameter sets. For RF number of input parameters selected randomly at each split – which is set under the ~~(~~'mtry' argument~~)~~ – was tuned. In the case of GBM influence of interaction depth and shrinkage were analysed. In ranger RF default value is 500 for the number of trees, that was used for both RF and GBM. Also for minimum number of observations in the terminal nodes of the trees the default value of the algorithms was used. ~~Optimization of input variable selection was performed based on variable importance calculated d~~During the tuning of model parameters the importance of variables was calculated both for GBM and RF methods to eliminate the less relevant predictors (Gregorutti et al., 2017; Nussbaum et al., 2018). Variable importance is the measure of relevance of each predictor, it is calculated from the average sum of squared improvements at each split, where the predictor was selected to partition the data (Hastie et al., 2009). A value of 100 is assigned to the largest variable importance value and the others are scaled accordingly to provide relative measure. The m~~M~~ost important 50-50 ~~independent variables~~predictors out of 173 for topsoils and 170 for subsoils have been selected from both GBM and RF models. After concatenating the 50-50 most important variables~~, then~~ parameter tuning was performed again with the decreased number of ~~input variables~~predictors.

We compared the accuracy of all models based on the cross-validation results and built the final prediction model (PTF) with the ~~best~~ better performing and ~~simplest~~ simpler algorithm on all training data with the optimized parameters. Performance of the PTFs was ~~described with root mean square error (~~RMSE~~)~~ (Eq. 1) and coefficient of determination (R^2) (Eq. 2).

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{\boldsymbol{MSE}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (2)$$

Performance of PTFs on the training dataset was ~~assumed~~ based on the results of a five-fold cross-validation, and out-of-bag samples – not included in the bootstrap sample used to build the tree – for GBM and RF respectively. In RF accuracy on out-of-bag samples was analysed. Uncertainty of the predictions was characterized with the 5 and 95% quantiles of the predicted values, calculated within the 'ranger' and 'gbm' packages during the ~~deriving~~ derivation of the prediction algorithms.

HUN-PTFs derived on the MARTHA dataset were used to calculate the soil hydraulic properties (THS, FC, WP) based on the available soil and environmental covariates available for the catchment (Table 1, section 2.1) as predictors, hence those were mapped indirectly. Soil information is currently available for the 0-30, 30-60 and 60-90 cm. The input information depth was set to 15, 45 and 75 cm for the first, second and third layer respectively during the calculation of soil hydraulic property maps.

We provided information on the uncertainty of the predictions by pixels ~~based.~~, ~~further~~ Further to the median and the~~value~~ 5 and 95% quantiles of the predicted values were also mapped for each soil hydraulic property. The prediction intervals were calculated by ~~based on~~ the PTFs.

2.4.2 Direct mapping with geostatistical method (RFK)

We applied random forest combined with kriging (RFK), which can be considered as a new 'workhorse' of digital soil mapping (Keskin and Grunwald, 2018). In the case of RFK, the deterministic component of spatial soil variation is modelled by the RF introduced above, whereas the stochastic part of variation is modelled by kriging using the derived residuals.

For the geostatistical ~~geostatsitical~~ analysis those samples of the MARTHA database were selected which fall within the catchment plus a 5 km buffer zone area. The buffer zone was used to increase the accuracy of geostatistical calculations also at the border of the catchment. On the study site data of 359 soil profiles are available from the MARTHA (Fig. 3~~2~~). Table 2 ~~shows~~ summarizes the measured soil chemical, physical, hydraulic data of the soil profiles' horizons.

First of all, we harmonized the soil hydraulic dataset for the required soil depths (i.e. 0-30, 30-60, 60-90 cm) by using equal-area splines (Malone et al., 2009), then we used RFK for predicting each soil hydraulic property for each soil depth, respectively. For RF we also optimized the parameter set by the 'train' function of R package 'caret' using five times repeated five-fold cross-validation. The most important 50 covariates – out of 173 for topsoils and 170 for subsoils,~~—~~ listed in Table 1 – have been selected and the final RF model was optimized with those ~~covariates~~predictors. We used the final RF model for predicting the deterministic component. We computed the residuals and then we estimated their variogram by Matheron's (1963) method-of-moments estimator. An isotropic variogram model was fitted to the estimated variogram by the 'fit.variogram' function of R package 'gstat' (Gräler et al., 2016; Pebesma, 2004). We kriged the residuals and then we added

~~back~~ them to the deterministic component predicted by RF. The above described modelling procedure was applied for each soil hydraulic property and for each soil depth. Performance of RF was described with RMSE (Eq. 1) and R^2 (Eq. 2).

2.4.3 Evaluating the performance of soil hydraulic maps

Performance of soil hydraulic maps was evaluated based on ~~measured~~ observed soil hydraulic properties ~~calculated~~ harmonized for 0-30, 30-60 and 60-90 cm depth with the method described in 2.4.2 section. RMSE and mean square error skill score (SS_{mse}) (Nussbaum et al., 2018) Eq. (1-3) were calculated for each map~~s~~.

$$SS_{mse} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}\left(y_i - \frac{1}{N}\sum_{i=1}^{N} y_i\right)^2} \quad (3)$$

Performance of soil hydraulic maps derived with HUN-PTFs and RFK was compared to the 3D European soil hydraulic maps (EU-SoilHydroGrids v1.0) (Tóth et al., 2017) ~~(EU-SHG)~~. In EU-SoilHydroGrids input information for mapping was SoilGrids 250 m (Hengl et al., 2017) on which EU-PTFs (Tóth et al., 2015) were applied, hence its resolution is 250 m. We converted the information of EU-SoilHydroGrids ~~SHG~~ to 0-30, 30-60 and 60-90 cm to be able to compare its performance to the 100 m resolution new soil hydraulic maps derived by HUN-PTFs and RFK.

~~For the comparison of the PTFs with different input variables and then the soil hydraulic maps derived with different methods~~ The Kruskal Wallis test implemented in the R package 'agricolae' (De Mendiburu, 2017) was applied at 5% significance level on the mean squared error values for the comparison of the PTFs with different input variables and also the soil hydraulic maps derived using different methods.

All statistical analyses were performed in R (R Core Team, 2017) ~~(R Core Team, 2017)~~.

3 Results and discussion

3.1 Pedotransfer functions

During the parameter tuning of RF and GBM we found that decreasing number of input variables – from 173 to 69-76 and from 170 to 65-77 in case of topsoil and subsoil predictions respectively – significantly improved prediction of top- and subsoil FC and subsoil WP. Although differences between RMSE values were less than 0.0001 cm^3 cm^{-3}, ~~which is~~these are negligible from a practical point of view. In Nussbaum et al. (2018) the number of input parameters were decreased from 300-500 environmental covariates to the 10, 20, 30, 40, 50 most important ones. No changes in performance were found~~They didn't find any change in performance~~ during validation. We can assume that performance of predictions will neither increase nor decrease if ~~only~~ more important independent variables are used exclusively for the predictions. Although the selection of the

most important independent variables can reduce (i) unnecessarily large size of the model which can speed up mapping of soil hydraulic properties for larger areas at fine resolution and (ii) multicollinearity between predictor variables. ~~PTFs derived with RF are not sensible for reducing independent variables to the most important ones,~~ Dorman et al. (2013) extensively studied the problem of collinearity to test its impact on predictions of ecological parameters. They analysed multiple regression and machine-learning methods and found that prediction performance of random forest did not get worse due to high collinearity in the training dataset even when structure of collinearity was different in training and validation data.~~neither multicollinearities between independent variables decrease performance~~. Influence of multicollinearity on the prediction performance is partly reduced due to the random selection features of RF but could be further elaborated in the presented methods, however this was beyond the scope of the presented work. ~~Although selection of most important independent variables can reduce unnecessarily large size of the model which can speed up mapping of soil hydraulic properties for larger areas at fine resolution.~~

In the case of RF optimal number of input parameters randomly selected at each split was between 10 and 20, depending on soil hydraulic parameter. In GBM optimal interaction depth varied between 20 and 40. Iteration converged during the prediction of lower 5% and upper 95% quantiles, but did not for 50%, which is the most probable predicted value. Therefore, the influence of shrinkage and increasing the number of trees to 1,000 was also analysed ~~as well~~ but only in the prediction of FC because training with low shrinkage values is very time consuming. We tuned shrinkage 0.1 and 0.01 with both 500 and 1,000 trees, setting interaction depth to 4, 6 and 10. Shrinkage with 0.1 value was more accurate than 0.01 independently from the number of trees and increasing number of trees did not significantly improve the prediction, therefore shrinkage was set to 0.1 and the default 500 number of trees were used in the algorithm.

Performance of PTFs derived by RF and GBM on training and test sets is included in Table 3. In the case of all soil hydraulic properties RF performed significantly better than GBM based on MSE on TEST and TEST_CHEM sets both for topsoil and subsoil predictions, except for WP topsoil predictions, where there was no significant difference between the methods. In this way PTFs derived with RF method were selected for mapping soil hydraulic properties. RMSE values calculated on the test sets for RF were between 0.042~~0~~ and 0.045~~3~~ cm^3 cm^{-3} for THS, 0.039~~40~~ and 0.042 cm^3 cm^{-3} for FC, 0.035~~6~~ and 0.038~~9~~ cm^3 cm^{-3} for WP, which is close to the performance of other internationally accepted PTFs (e.g. Botula et al. (2013), Román Dobarco et al. (2019), Zhang and Schaap (2017)). R^2 was 0.408~~0~~-0.487, 0.~~739~~746-0.~~770~~ 766 and 0.~~711~~737-0.762 for THS, FC and WP respectively on test sets in the case of RF. Figure 4. shows the scatterplots of measured versus predicted values with the 90% prediction interval. At the lower end of the soil hydraulic property distribution, real values were closer to the lower 5% quantile predictions, at the higher end of its distribution the real values are closer to the upper 95% quantile predictions. When we compared performance of RF derived for topsoils – which includes organic matter content~~OM~~, pH and calcium carbonate content~~CaCO~~, as well among the input parameters - and subsoils there was no significant difference based on the results in the TEST_CHEM set. This is due to their correlation with other environmental ~~covariates~~ predictors considered in the PTFs such as soil texture, depth, longitude, elevation, slope angle, multi-resolution valley bottom flatness, horizontal distance to existing water bodies, roughness, temperature, precipitation, solar radiance, spectral reflectance in red and near

infrared and normalized difference vegetation index (Adhikari et al., 2014; Hengl et al., 2017; Nussbaum et al., 2018). When other environmental covariates than soil related variables are not included among input parameters chemical properties significantly improve prediction (Hodnett and Tomasella, 2002; Khodaverdiloo et al., 2011; Tóth et al., 2015). In the case of THS range, the ~~of~~ predicted values using chemical parameters as well were closer to the range of measured values, therefore we also considered soil chemical properties ~~as well~~ for the topsoil predictions. For FC and WP range of values predicted with PTF not including chemical variables were closer to that of measured values, hence information on organic matter content~~OM~~, pH and calcium carbonate content~~$CaCO_3$~~ – even though it is available – was not considered during the estimation of topsoil hydraulic properties.

The presented PTFs were derived on the full MARTHA dataset, therefore those are applicable to predict the THS, FC and WP of soils in the whole Pannonian region.

3.1.1 Importance of independent variables

For THS organic matter content~~OM~~, silt, sand content, pH, clay, calcium carbonate~~$CaCO_3$~~ content are the most important variables with relative importance of over 20% based on final RF model. Further to those properties, soil depth, mean annual precipitation, mean monthly maximum, minimum and mean temperature of some months, mean monthly radiation, longitude, horizontal and vertical distance to existing water bodies, multi-resolution valley bottom flatness and ridge top flatness, water vapour pressure in August, spectral reflectance in near infrared are among the most important 30 variables having 10-15 % relative importance. For FC and WP clay, silt and sand content~~PSD~~ and ~~OM~~ organic matter content are the most important variables, having relative importance around and over 20 %. ~~ST~~Soil type, mean monthly precipitation in July, vertical distance to existing water bodies and longitude have relative importance around 5-14 % in case of FC. All the other environmental covariates have relative importance of less than 5%. For WP longitude, mean monthly precipitation of November and July, elevation, vertical and horizontal distance to existing water bodies, calcium carbonate content~~$CaCO_3$~~, mean monthly radiation, pH, depth, mean monthly water vapour pressure, multi-resolution ridge top flatness and spectral reflectance in near infrared have relative importance of between 5-16 %. Information on topography was found important for the prediction of soil hydraulic properties by Obi et al. (2014), Rawls and Pachepsky (2002), Romano and Chirico (2004), Zhao et al. (2016) as well. Information on land cover was not retained after selecting the most important variables.

When soil chemical properties (organic matter content~~OM~~, calcium carbonate content~~$CaCO_3$~~, pH) are not included among input parameters, sand, silt, clay content are far the most important three independent variables (39-100 %). In the case of THS also depth has higher relative importance (52 %). For the prediction of FC importance of ~~ST~~ soil type increases to 18 %. In ~~case of~~ WP prediction there is no notable change in variable importance when chemical properties are not included in the RF.

Summary of the variable importance analysis showed that soil properties are far the most important input parameters for the prediction of soil hydraulic properties (Fig. 5~~3~~). In this way resolution of soil maps determined the resolution of the derived soil hydraulic maps, which was 100 m.

3.2 Random forest combined with kriging (RFK)

During the RF parameter tuning we also found that decreasing the number of environmental covariates – from 173 to 50 and from 170 to 50 in the case of topsoil and subsoil~~,~~ respectively – significantly improved the prediction accuracy for each soil hydraulic property. For the final RF models the optimal number of randomly selected predictors at each split~~ntry values~~ varied between 5 and 40 depending on the given soil hydraulic property. The performance of the final RF models are summarized in Table 4. R^2 varies between 0.189-0.403, 0.478-0.562 and 0.463-474 for THS, FC and WP, respectively. RMSE was 0.055-0.060, 0.053-0.063 and 0.051-0.056 for THS, FC and WP, respectively. For describing spatial variation of the soil hydraulic properties the most important environmental covariates were the ~~ST~~soil type, ~~OM~~ organic matter content (for topsoil), clay, silt and sand content and the pH (for topsoil). The final RF models were used for estimating the deterministic component for each soil hydraulic property.

The parameters of the fitted variogram models are summarized in Table 4. In the case of exploratory variography most of the experimental variograms did not show spatial structure and the applied variogram fitting algorithm was not able to find a satisfactory variogram model in case of six out of nine under 200 iterations. Hence, a nugget model was fitted to those variograms (Table 4), which is not rare in digital soil mapping (Hengl et al., 2015; Szatmári and Pásztor, 2018; Vaysse and Lagacherie, 2017). In Table 4 we have observed that the lower the R^2 value was, the higher the range parameter became. ~~Furthermore, there is a relationship between the R^2 and range values (see Table 4), i.e. the lower the R^2 the higher the range parameter.~~ The fitted variogram models were used for kriging of the RF residuals for each soil hydraulic property. We summed the RF predictions and the kriged residuals to get the RFK maps for each of the target hydraulic properties.

3.3 Performance of soil hydraulic maps

formázott: Címsor 1

New 100 m resolution soil hydraulic maps significantly outperformed the EU-SoilHydroGrids (Table 5), which was expected because (i) reference soil data originate from the mapped area, (ii) also spatially denser and (iii) locally trained models are used. ~~Further to it~~In addition, several environmental covariates were considered for the predictions and relationship between easily available soil properties and soil hydraulic parameters were derived ~~on~~ from local data.

In the case of mapping six out of nine soil hydraulic maps there was no significant difference between maps derived by RFK and HUN-PTFs. In the case of THS HUN-PTF performed significantly better for mapping the 30-60 and 60-90 cm. For calculating WP at 60-90 cm soil depth RFK was significantly better than HUN-PTF method.

The ~~R~~range of predicted values is smaller in the case of HUN-PTF method than in RFK, which is due to the "averaging approach" of the algorithm which in the case of RFK is spatially corrected allowing a wider range in the predicted values (Fig. 6, 7, 8, 9~~4, 5~~). Density plot of predicted values are smoother in the case of RFK than in HUN-PTF and EU-SoilHydroGrids~~SHG~~ maps (Fig. 6~~4~~). This is due to adding residuals of kriging which modifies the values derived by random forest. In EU-SoilHydroGrids~~SHG~~ soil hydraulic values were calculated with linear regression based on soil properties available from SoilGrids, where mapping was performed with RF without kriging. In this way possible soil input combinations are limited in the ~~EU-SHG~~European maps. In SoilGrids algorithms are derived ~~on~~ from a global dataset (Hengl et al., 2017), which has sparser measured data than the Hungarian soil profile database used to map soil properties (Laborczi et al., 2018; Szatmári and Pásztor, 2018). ~~Further to it~~In addition, RF is based on an averaging algorithm, ~~these~~ which limits the ability to describe local extreme values. These result in smaller range and variability of calculated soil hydraulic properties on EU-SoilHydroGrids~~SHG~~ maps than on RFK or HUN-PTF ones (Fig. 4~~6~~) The basic Hungarian soil maps were derived with regression kriging methods, thus providing~~e~~ smoother soil input data for the calculations. As an example of how differences in the range of predicted soil hydraulic properties can be visualized, the maps of THS, FC, WP are~~is~~ shown on Fig. ~~5~~7-9 (a), (b), (c) for a selected area of the catchment. Differences between the new and already available maps also comes ~~also from~~ occur due to the differences in resolution, which is 100 m for RFK and HUN-PTF and 250 m for EU- SoilHydroGrids~~SHG~~. Even though the influence of topographical information was less than that of soil properties when PTFs were derived, the pattern of topography is visible on the maps derived by RFK and HUN-PTFs. This is due to the soil layers used as inputs for calculating the soil hydraulic properties, because topographical information was important among the covariates when the maps on them were derived (Szatmári et al., 2013). In RFK influence of the topography is less visible, it could be smoothed by adding kriged residuals. A ~~M~~map of possible lower 5 % and upper 95 % values based on the HUN-PTF method are also shown in Fig. ~~5~~7-9 (d), (e). The range between the lower and upper possible values (Fig. 10~~6~~) are usually higher for Histosols, Gleysols and Luvisols under forest land use, because these kind of soils are underrepresented in the MARTHA database.

In this study our aim was to analyse performance of the PTF approach and machine learning algorithm~~MLA~~ combined with geostatistics to derive soil hydraulic maps and compare their accuracy to that of EU-SoilHydroGrids. Therefore we ~~didn't~~ have not differentiated uncertainty of the maps originating from the soil input layers – i.e. DOSoReMI.hu and SoilGrids.

Average difference between the RFK and HUN-PTFs maps is between 0.003 and 0.012 $cm^3\ cm^{-3}$ for THS, 0.011 and 0.015 $cm^3\ cm^{-3}$ for FC, 0.015 and 0.018 $cm^3\ cm^{-3}$ for WP, depending on soil depth. Absolute difference between the maps derived with HUN-PTFs and RFK is less than 0.025 $cm^3\ cm^{-3}$ for at least 65 % of the mapped area and was always smaller than 0.100 $cm^3\ cm^{-3}$ (Table 6). On those areas where difference between RFK and HUN-PTF was higher than 0.025 $cm^3\ cm^{-3}$, HUN-PTF predicted lower water retention at all matric potential values for Histosols and Luvisols under forest land use type. WP values predicted with HUN-PTFs were higher than that of RFK for Luvisols with sandy texture and under forest land use type.

Based on SS_{mse} values in the case of seven out of nine soil hydraulic maps RFK mapping method was more accurate than HUN-PTF, although only calculation of WP in 60-90 cm depth was significantly better. For THS HUN-PTFs performed significantly better at 30-60 and 60-90 cm soil depth.

In this study priority was put on the usability and transferability of the results into practical applications. The purpose of the presented research was to derive as accurate maps as possible. Thus ability for full comparability of the methods did not determine design of methodology and statistical analysis. Therefore, in the RFK analysis all measured data were used for the mapping. For the PTF approach predictions were tested on randomly selected 33% samples of the whole MARTHA database without distinguishing samples located on the catchment, as it is usually done in deriving PTFs. This provides broader information and possibility for a wider application of ~~the~~ PTFs. The presented HUN-PTF mapping method can be applied in any catchments of Hungary.

3.4 Practical use of the analysis

formázott: Címsor 1

RF performed significantly better than GBM in 7 cases out of 8 on test sets. RF was found to be a suitable method to provide information on the prediction uncertainty, any desired quantiles of the predicted value can be calculated. This enables it to include extreme soil hydraulic parameters for hydrological simulations. Its further advantage is that it can handle several independent variables, performance of prediction is not influenced by multicollinearity between independent variables and inclusion of ~~not~~ unimportant input parameter. Calculation on multiple cores is implemented in the random forest algorithm in 'ranger' R package, which can significantly decrease computation time.

Easily available soil properties such as sand, silt and clay content, organic matter content and depth were the most important input variables for the calculation of THS, FC and WP among the analysed 173 soil and environmental covariates. For THS calcium carbonate content and pH were also among independent variables with higher importance. Geographical coordinates, information on topography, climate and vegetation had smaller relative importance. Covariates on land use and parent material were not among the 50 most important variables. Therefore, resolution of available soil maps determined the resolution of new soil hydraulic maps, which is 100 m.

The number of input variables can be decreased based on variable importance, which can significantly decrease computation time and information not relevant for the prediction can be discarded. For practical application it is desirable to decrease the size of the prediction models when PTFs are applied for soil hydraulic mapping at country scale at finer resolution.

If data on topography, climate and vegetation are also considered for the prediction missing information on chemical properties, such as organic matter content~~OM~~, pH, calcium carbonate content~~$CaCO_3$~~ can be covered by the environmental covariates without significant loss of performance.

HUN-PTFs performed significantly better for the prediction of THS at 30-60 and 60-90 cm depth, although the absolute difference between the RFK and HUN-PTFs maps ~~are~~is less than 0.025 cm^3 cm^{-3} for at least 75 % of the area. Spatial patterns

of topography ~~is~~are less dominant on the soil hydraulic maps prepared by the RFK method due to kriging the residuals, which is an advantage. Maps prepared by the HUN-PTFs cannot decrease the influence of topography included in the input layers therefore even if topographical parameters are not important for the prediction of soil hydraulic properties ~~those~~that are visible on the soil hydraulic maps. Considering all these results we suggest ~~to use~~using the soil hydraulic maps prepared by the RFK only if ~~only~~ the most probable soil hydraulic value is needed for the Balaton catchment area. Information on the uncertainty of the predicted values can be derived with geostatistical methods as well, e.g. Szatmári and Pásztor (2018), Rudiyanto et al. (2016), Viscarra Rossel et al. (2015) presented possible methods. ~~Although deriving it with RFK is time consuming, labour and computation intensive.~~According to Szatmári and Pásztor (2018), 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 requiring 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. If information on uncertainty is needed as well maps derived by the HUN-PTFs are recommended to use. In Table 7 we highlighted the most important differences between pedotransfer function (HUN-PTF) and geostatistical (RFK) based soil hydraulic mapping based on the Balaton catchment. Most of the findings are in line with Hengl et al. (2018b), Tranter et al. (2009), Vaysse and Lagacherie (2017), Webster and Oliver (2007).

formázott: Listaszerű bekezdés

4 Conclusions

~~Our aim was to analyse performance of different soil hydraulic mapping methods for the Balaton catchment in Hungary. We mapped soil hydraulic properties at 100 m resolution with i) applying pedotransfer function derived on a country wide soil hydraulic dataset (HUN-PTFs), ii) geostatistical method using random forest and kriging (RFK) based on environmental covariates available for the catchment. Then we compared their performance to the 250 m resolution 3D European soil hydraulic maps (EU-SoilHydroGrids).~~Based on results of six out of nine soil hydraulic maps there is no significant difference in performance between values derived using pedotransfer function and geostatistical method on the Balaton catchment area. The benefit of maps computed with random forest and kriging is that locally extreme values can be characterized better. 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 analyse the difference between uncertainty maps calculated with the different methods specifically for soil hydraulic properties.

~~Easily available soil properties such as sand, silt and clay content, OM and depth were the most important input variables for the calculation of THS, FC and WP among the analysed 173 soil and environmental covariates. For THS $CaCO_3$ and pH were also among independent variables with higher importance. Geographical coordinates, information on topography, climate and~~

~~vegetation had smaller relative importance. Covariates on land use and parent material were not among the 50 most important variables. Therefore, resolution of available soil maps determined the resolution of new soil hydraulic maps, which is 100 m. Number of input variables can be decreased based on variable importance, which can significantly decrease computation time and information not relevant for the prediction can be discarded. For practical application it is desirable to decrease size of the prediction models when PTFs are applied for soil hydraulic mapping at country scale at finer resolution.~~

~~RF performed significantly better than GBM in 7 cases out of 8 on test sets. RF was found to be a suitable method to provide information on the prediction uncertainty, any desired quantiles of the predicted value can be calculated. This enables to include extreme soil hydraulic parameters for hydrological simulations. Its further advantage that it can handle several independent variables, performance of prediction is not influenced by multicollinearity between independent variables and inclusion of not important input parameter. Calculation on multiple cores is implemented in the random forest algorithm in 'ranger' R package, which can significantly decrease computation time.~~

~~If data on topography, climate and vegetation are also considered for the prediction missing information on chemical properties, such as OM, pH, $CaCO_3$ can be covered by the environmental covariates without significant loss of performance.~~

~~HUN-PTFs performed significantly better for the prediction of THS at 30-60 and 60-90 cm depth, although absolute difference between the RFK and HUN-PTFs maps are less than 0.025 $cm^3 cm^{-3}$ for at least 75 % of the area. Spatial patterns of topography is less dominant on the soil hydraulic maps prepared by the RFK method due to kriging the residuals, which is an advantage. Maps prepared by the HUN-PTFs cannot decrease the influence of topography included in the input layers therefore even if topographical parameters are not important for the prediction of soil hydraulic properties those are visible on the soil hydraulic maps. Considering all these results we suggest to use the soil hydraulic maps prepared by the RFK if only the most probable soil hydraulic value is needed for the Balaton catchment. Information on uncertainty of the predicted values can be derived with geostatistical methods as well, e.g. Szatmári and Pásztor (2018), Rudiyanto et al. (2016), Viscarra Rossel et al. (2015) presented possible methods. Although deriving it with RFK is time consuming, labour and computation intensive. If information on uncertainty is needed as well maps derived by the HUN-PTFs are recommended to use. In Table 7 we highlighted the most important differences between pedotransfer function (HUN-PTF) and geostatistical (RFK) based soil hydraulic mapping based on the Balaton catchment. Most of the findings are in line with Hengl et al. (2018b), Tranter et al. (2009), Vaysse and Lagacherie (2017), Webster and Oliver (2007).~~

Data availability. The 3D soil hydraulic maps of the Balaton catchment – in GeoTIFF format – and the hydraulic pedotransfer functions – in RData format – are freely available for non-commercial use from the Institute for Soil Sciences and Agricultural

formázott: Címsor 1, Sorköz: 1,5 sor

Chemistry Centre for Agricultural Research Hungarian Academy of Sciences (http://mta-taki.hu/en/kh124765/maps~~),~~ https://www.mta-taki.hu/en/kh124765/hun_ptfs ~~in TIFF format~~).

Author contribution. BT conceptualized the study, designed the methodology and coordinated the research. AM provided the MARTHA dataset. LP and AL cured soil maps, KT prepared all the other covariate layers. AM, LP, KT, AL, GSZ, BT performed data curation. GSZ carried out geostatistical analysis, BT derived the PTFs, they applied statistical and computational analysis. AL assisted in visualization of maps and built website for data download. KR, AM, LP contributed to the interpretation. LP provided the computing resources. BT prepared the paper with considerable input from GSZ and further contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The research project was supported by the Hungarian National Research, Development and Innovation Office (NRDI) under grants KH124765, KH126725, K119475, through the common grant of the Hungarian and Polish Academy of Sciences (Grant No. NKM-108/2017)" and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B. and Greve, M. H.: Digital mapping of soil organic carbon contents and stocks in Denmark, PLoS One, 9(8), doi:10.1371/journal.pone.0105519, 2014.

Ahuja, L. R., Naney, J. W. and Williams, R. D.: Estimating soil water characteristics from simpler properties or limited data., Soil Sci. Soc. Am. J., 49(5), 1100–1105, doi:https://doi.org/10.2136/sssaj1985.03615995004900050005x, 1985.

Baker, L. and Ellison, D.: Optimisation of pedotransfer functions using an artificial neural network ensemble method, Geoderma, 144(1–2), 212–224, doi:10.1016/j.geoderma.2007.11.016, 2008.

Bashfield, A. and Keim, A.: Continent-wide DEM Creation for the European Union, in 34th International Symposium on Remote Sensing of Environment - The GEOSS Era: Towards Operational Environmental Monitoring. [online] Available from: http://www.isprs.org/proceedings/2011/isrse-34/211104015Final00143.pdf, 2011.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T. and MacMillan, R. A.: Spatial modelling with Euclidean distance fields and machine learning, Eur. J. Soil Sci., 69(5), 757–770, doi:10.1111/ejss.12687, 2018.

Botula, Y.-D., Nemes, A., Mafuka, P., Van Ranst, E. and Cornelis, W. M.: Prediction of Water Retention of Soils from the Humid Tropics by the Nonparametric -Nearest Neighbor Approach, Vadose Zo. J., 12(2), 1–17., doi:10.2136/vzj2012.0123, 2013.

Bouma, J.: Using Soil Survey Data for Quantitative Land Evaluation, pp. 177–213, Springer US., 1989.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A.: Classification and Regression Trees, Chapman and Hall/CRC. [online] Available from: http://www.amazon.com/Classification-Regression-Trees-Leo-Breiman/dp/0412048418 (Accessed 2 May 2013), 1984.

Caruana, R. and Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms, Int. Conf. Mach. Learn., 161–168, doi:10.1145/1143844.1143865, 2006.

Caruana, R., Karampatziakis, N. and Yessenalina, A.: An empirical evaluation of supervised learning in high dimensions, Proc. 25th Int. Conf. Mach. Learn. - ICML '08, 96–103, doi:10.1145/1390156.1390169, 2008.

CEC EEA: CORINE land cover, , http://land.copernicus.eu/pan-european/corine-land, 2012.

Chaney, N. W., Wood, E. F., McBratney, A. B., Hempel, J. W., Nauman, T. W., Brungard, C. W. and Odgers, N. P.: POLARIS: A 30-meter probabilistic soil series map of the contiguous United States, Geoderma, 274, 54–67, doi:10.1016/j.geoderma.2016.03.025, 2016.

Chen, S., Richer-de-Forges, A. C., Saby, N. P. A., Martin, M. P., Walter, C. and Arrouays, D.: Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area, Geoderma, 312(June 2017), 52–63, doi:10.1016/j.geoderma.2017.10.009, 2018.

Cichota, R., Vogeler, I., Snow, V. O. and Webb, T. H.: Ensemble pedotransfer functions to derive hydraulic properties for New Zealand soils, Soil Res., 51(2), 94–111, doi:10.1071/SR12338, 2013.

Cisty, M., Celar, L. and Minaric, P.: Conversion between soil texture classification systems using the random forest algorithm, Air, Soil Water Res., 8, 67–75, doi:10.4137/ASWR.S31924, 2015.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V. and Böhner, J.: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model Dev., 8, 1991–2007, doi:10.5194/gmd-8-1991-2015, 2015.

Dai, Y., Shangguan, W., Duan, Q., Liu, B., Fu, S. and Niu, G.-Y.: Development of a China Dataset of Soil Hydraulic Parameters Using Pedotransfer Functions for Land Surface Modeling, J. Hydrometeorol., 14(3), 869–887, doi:10.1175/JHM-D-12-0149.1, 2013.

Dharumarajan, S., Hegde, R. and Singh, S. K.: Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India, Geoderma Reg., 10(April), 154–162, doi:10.1016/j.geodrs.2017.07.005, 2017.

Dietterich, T. G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees, Mach. Learn., 40, 139–157, doi:10.1023/A:1007607513941, 2000.

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.

Farkas, C., Rajkai, K., Kertész, M., Bakacsi, Z. and Meirvenne, M.: Spatial variability of soil hydro-physical properties: A case study in Herceghalom, Hungary., in Soil geography and geostatistics. Concepts and Applications, edited by P. Krasilnikov, F. Carré, and L. Montanarella, pp. 107–128, Joint Research Centre, Luxembourg. [online] Available from: https://esdac.jrc.ec.europa.eu/ESDB_Archive/eusoils_docs/other/EUR23290.pdf, 2008.

Ferrer Julià, M., Estrela Monreal, T., Sánchez Del Corral Jiménez, A. and García Meléndez, E.: Constructing a saturated hydraulic conductivity map of Spain using pedotransfer functions and spatial prediction, Geoderma, 123(3–4), 257–277, doi:10.1016/j.geoderma.2004.02.011, 2004.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, Int. J. Climatol., 37(12), 4302–4315, doi:10.1002/joc.5086, 2017.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine, Ann. Stat., 29(5), 1189–1232, doi:DOI 10.1214/aos/1013203451, 2001.

Gräler, B., Pebesma, E. J. and Heuvelink, G. B. M.: Spatio-Temporal Interpolation using gstat, R J., 8(1), 204–218, 2016.

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.

Gyalog, L. and Síkhegyi, F.: Magyarország földtani térképe, M=1:100 000 (Geological map of Hungary, M=1:100,000), Magyar Állami Földtani Intézet, Budapest. [online] Available from: https://map.mfgi.hu/fdt100/, 2005.

Hastie, T., Tibshirani, R. and Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction., Second edi., Springer. [online] Available from: https://web.stanford.edu/~hastie/ElemStatLearn//printings/ESLII_print10.pdf (Accessed 19 November 2018), 2009.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., De Jesus, J. M., Tamene, L. and Tondoh, J. E.: Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions, PLoS One, 10(6), 1–26, doi:10.1371/journal.pone.0125814, 2015.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, edited by B. Bond-Lamberty, PLoS One, 12(2), e0169748, doi:10.1371/journal.pone.0169748, 2017.

Hengl, T., Walsh, M. G., Sanderman, J., Wheeler, I., Harrison, S. P. and Prentice, I. C.: Global mapping of potential natural vegetation: an assessment of Machine Learning algorithms for estimating land potential, PeerJ Prepr., (73), doi:10.7287/peerj.preprints.26811v1, 2018a.

Hengl, T., Nussbaum, M., Wright, M. N. and Heuvelink, B. M.: Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-temporal Variables, , (May), doi:10.7287/peerj.preprints.26693v3, 2018b.

Hodnett, M. G. and Tomasella, J.: Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: a new water-retention pedo-transfer functions developed for tropical soils, Geoderma, 108(3–4), 155–180, doi:10.1016/S0016-7061(02)00105-2, 2002.

IUSS Working Group WRB: World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps., Rome., 2014.

Keskin, H. and Grunwald, S.: Regression kriging as a workhorse in the digital soil mapper's toolbox, Geoderma, 326(March), 22–41, doi:10.1016/j.geoderma.2018.04.004, 2018.

Khodaverdiloo, H., Homaee, M., van Genuchten, M. T. and Dashtaki, S. G.: Deriving and validating pedotransfer functions for some calcareous soils, J. Hydrol., 399(1–2), 93–99, doi:10.1016/j.jhydrol.2010.12.040, 2011.

Kishné, A. S., Tadesse, Y., Morgan, C. L. S. and Dornblaser, B. C.: Evaluation and improvement of the default soil hydraulic parameters for the Noah Land Surface Model, Geoderma, 285, 247–259, doi:10.1016/j.geoderma.2016.09.022, 2017.

Koestel, J. and Jorda, H.: What determines the strength of preferential transport in undisturbed soil under steady-state flow?, Geoderma, 217–218, 144–160, doi:10.1016/j.geoderma.2013.11.009, 2014.

Kuhn, M., Wing, J., Weston, S., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R. C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt, T.: caret: Classification and Regression Training. R package version 6.0-79. [online] Available from: https://github.com/topepo/caret/, 2018.

Laborczi, A., Szatmári, G., Kaposi, A. D. and Pásztor, L.: Comparison of soil texture maps synthetized from standard depth layers with directly compiled products, Geoderma, (January), doi:10.1016/j.geoderma.2018.01.020, 2018.

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.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., Schaap, M. G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M. J., Weihermüller, L., Zacharias, S., Zhang, Y. and Vereecken, H.: Pedotransfer Functions in Earth System Science: Challenges and Perspectives, Rev. Geophys., 55(4), 1199–1256, doi:10.1002/2017RG000581, 2017.

Makó, A., Tóth, B., Hernádi, H., Farkas, C. and Marth, P.: Introduction of the Hungarian Detailed Soil Hydrophysical Database (MARTHA) and its use to test external pedotransfer functions, Agrokémia és Talajt., 59(1), 29–38, 2010.

Malone, B. P., McBratney, A. B., Minasny, B. and Laslett, G. M.: Mapping continuous depth functions of soil carbon storage and available water capacity, Geoderma, 154(1–2), 138–152, doi:10.1016/j.geoderma.2009.10.007, 2009.

Marthews, T. R., Quesada, C. A., Galbraith, D. R., Malhi, Y., Mullins, C. E., Hodnett, M. G. and Dharssi, I.: High-resolution hydraulic parameter maps for surface soils in tropical South America, Geosci. Model Dev., 7(3), 711–723, doi:10.5194/gmd-7-711-2014, 2014.

Matheron, G.: Principles of geostatistics, Econ. Geol., 58(8), doi:10.2113/gsecongeo.58.8.1246, 1963.

Matos-Moreira, M., Lemercier, B., Dupas, R., Michot, D., Viaud, V., Akkal-Corfini, N., Louis, B. and Gascuel-Odoux, C.: High-resolution mapping of soil phosphorus concentration in agricultural landscapes with readily available or detailed survey data, Eur. J. Soil Sci., 68(3), 281–294, doi:10.1111/ejss.12420, 2017.

McNeill, S. J., Lilburne, L. R., Carrick, S., Webb, T. H. and Cuthill, T.: Pedotransfer functions for the soil water characteristics

of New Zealand soils using S-map information, Geoderma, 326(April), 96–110, doi:10.1016/j.geoderma.2018.04.011, 2018.

De Mendiburu, F.: Statistical Procedures for Agricultural Research, , 680, doi:10.2307/2411227>., 2017.

Montzka, C., Herbst, M., Weihermüller, L., Verhoef, A. and Vereecken, H.: A global data set of soil hydraulic properties and sub-grid variability of soil water retention and hydraulic conductivity curves, Earth Syst. Sci. Data Discuss., in review, doi:10.5194/essd-2017-13, 2017.

Motaghian, H. R. and Mohammadi, J.: Spatial Estimation of Saturated Hydraulic Conductivity from Terrain Attributes Using Regression, Kriging, and Artificial Neural Networks, Pedosphere, 21(2), 170–177, doi:10.1016/S1002-0160(11)60115-X, 2011.

Natekin, A. and Knoll, A.: Gradient boosting machines, a tutorial, Front. Neurorobot., 7(DEC), doi:10.3389/fnbot.2013.00021, 2013.

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.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E. and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL, 4, 1–22, doi:10.5194/soil-4-1-2018, 2018.

Obi, J. C., Ogban, P. I., Ituen, U. J. and Udoh, B. T.: Catena Development of pedotransfer functions for coastal plain soils using terrain attributes, Catena, 123, 252–262, doi:10.1016/j.catena.2014.08.015, 2014.

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A. and Moore, J. H.: Data-driven Advice for Applying Machine Learning to Bioinformatics Problems, , doi:10.1142/9789813235533_0018, 2017.

Pachepsky, Y., Shcherbakov, R., Várallyay, G. and Rajkai, K.: Soil water retention as related to other soil physical properties., Pochvovedenie, 2, 42–52, 1982.

Pásztor, L., Laborczi, A., Takács, K., Szatmári, G., Fodor, N., Illés, G., Farkas-Iványi, K., Bakacsi, Z. and Szabó, J.: Compilation of Functional Soil Maps for the Support of Spatial Planning and Land Management in Hungary, in Soil Mapping and Process Modeling for Sustainable Land Use Management, edited by P. Pereira, E. C. Brevik, M. Munoz-Rojas, and B. A. Miller, pp. 293–317, Elsevier, Amsterdam., 2017.

Pásztor, L., Laborczi, A., Bakacsi, Z., Szabó, J. and Illés, G.: Compilation of a national soil-type map for Hungary by sequential classification methods, Geoderma, 311(April 2017), 93–108, doi:10.1016/j.geoderma.2017.04.018, 2018a.

Pásztor, L., Laborczi, A., Takács, K., Szatmári, G., Bakacsi, Z., Szabó, J. and Illés, G.: DOSoReMI as the national implementation of GlobalSoilMap for the territory of Hungary, in Proceedings of the Global Soil Map 2017 Conference, July 4-6, 2017, edited by D. Arrouay, I. Savin, J. Leenaars, and A. B. McBratney, pp. 17–22, CRC Press, Moscow, Russia., 2018b.

Pebesma, E. J.: Multivariable geostatistics in S: The gstat package, Comput. Geosci., 30(7), 683–691, doi:10.1016/j.cageo.2004.03.012, 2004.

R Core Team: R: A language and environment for statistical computing., [online] Available from: https://www.r-project.org,

2017.

Ramcharan, A., Hengl, T., Beaudette, D. and Wills, S.: A Soil Bulk Density Pedotransfer Function Based on Machine Learning: A Case Study with the NCSS Soil Characterization Database, Soil Sci. Soc. Am. J., 81(0), 1279–1287, doi:10.2136/sssaj2016.12.0421, 2017.

Rawls, W. and Brakensiek, D.: Estimating soil water retention from soil properties., J. Irrig. Drain. Div., 108(2), 166–171, 1982.

Rawls, W. J. and Pachepsky, Y. A.: Using field topographic descriptors to estimate soil water retention., Soil Sci., 167(7), 423–435., 2002.

Ridgeway, G.: gbm: Generalized Boosted Regression Models. R package version 2.1.3., 2017.

Román Dobarco, M., Cousin, I., Le Bas, C. and Martin, M. P.: Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty, Geoderma, 336(April 2018), 81–95, doi:10.1016/J.GEODERMA.2018.08.022, 2019.

Romano, N. and Chirico, G. B.: The role of terrain analysis in using and developing pedotransfer functions, in DEVELOPMENTS IN SOIL SCIENCE, vol. 30, edited by Y. Pachepsky and W. J. Rawls, pp. 273–294, Elsevier, Amsterdam., 2004.

Rudiyanto, Minasny, B., Setiawan, B. I., Arif, C., Saptomo, S. K. and Chadirin, Y.: Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands, Geoderma, 272, 20–31, doi:10.1016/j.geoderma.2016.02.026, 2016.

Rudiyanto, Minasny, B., Setiawan, B. I., Saptomo, S. K. and McBratney, A. B.: Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands, Geoderma, 313(November 2017), 25–40, doi:10.1016/j.geoderma.2017.10.018, 2018.

Saxton, K. E., Rawls, W., Romberger, J. S. and Papendick, R. I.: Estimating generalized soil-water characteristics from texture, Soil Sci. Soc. Am. J., 50(4), 1031–1036, doi:https://doi.org/10.2136/sssaj1986.03615995005000040039x, 1986.

Sequeira, C. H., Wills, S. A., Seybold, C. A. and West, L. T.: Predicting soil bulk density for incomplete databases, Geoderma, 213, 64–73, 2014.

Souza, E. De, Batjes, N. H. and Pontes, L. M.: Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin, Sci. Agric., 73(6), 525–534, doi:10.1590/0103-9016-2015-0485, 2016.

Szatmári, G. and Pásztor, L.: Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms, Geoderma, 1–12, doi:10.1016/j.geoderma.2018.09.008, 2018.

Szatmári, G., Laborczi, A., Illés, G. and Pásztor, L.: Large-scale mapping of soil organic matter content by regression kriging in Zala County, Agrokémia és Talajt., 62(2), 219–234, doi:10.1556/Agrokem.62.2013.2.4, 2013.

Szentimrey, T. and Bihari, Z.: Mathematical background of the spatial interpolation methods and the software MISH (Meteorological Interpolation based on Surface Homogenized Data Basis)., in Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology., pp. 17–27, Budapest., 2007.

Tóth, B., Makó, A. and Tóth, G.: Role of soil properties in water retention characteristics of main Hungarian soil types, J. Cent. Eur. Agric., 15(2), 137–153, doi:10.5513/JCEA01/15.2.1465, 2014.

Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G. and Tóth, G.: New generation of hydraulic pedotransfer functions for Europe., Eur. J. Soil Sci., 66(1), 226–238, doi:10.1111/ejss.12192, 2015.

Tóth, B., Weynants, M., Pásztor, L. and Hengl, T.: 3D soil hydraulic database of Europe at 250 m resolution, Hydrol. Process., 31(14), 2662–2666, doi:10.1002/hyp.11203, 2017.

Tranter, G., McBratney, a. B. and Minasny, B.: Using distance metrics to determine the appropriate domain of pedotransfer function predictions, Geoderma, 149(3–4), 421–425, doi:10.1016/j.geoderma.2009.01.006, 2009.

Vaysse, K. and Lagacherie, P.: Using quantile regression forest to estimate uncertainty of digital soil mapping products, Geoderma, 291, 55–64, doi:10.1016/j.geoderma.2016.12.017, 2017.

Vereecken, H., Maes, J., Feyen, J. and Darius, P.: Estimating the Soil Moisture Retention Characteristic From Texture, Bulk Density, and Carbon Content, Soil Sci., 148(6), 389–403, doi:10.1097/00010694-198912000-00001, 1989.

Vermote, E.: MOD09A1 MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V006, , doi:10.5067/MODIS/MOD09A1.006, 2015.

Viscarra Rossel, R. A., Chen, C., Grundy, M. J., Searle, R., Clifford, D. and Campbell, P. H.: The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project, Soil Res., 53(8), 845–864, doi:10.1071/SR14366, 2015.

Webster, R. (Richard) and Oliver, M. A.: Geostatistics for environmental scientists, Wiley. [online] Available from: https://www.wiley.com/en-us/Geostatistics+for+Environmental+Scientists%2C+2nd+Edition-p-9780470028582 (Accessed 16 October 2018), 2007.

Wu, X., Lu, G. and Wu, Z.: An Integration Approach for Mapping Field Capacity of China Based on Multi-Source Soil Datasets, , 10, 728, doi:10.3390/w10060728, 2018.

Xu, Z., Wang, X., Chai, J., Qin, Y. and Li, Y.: Simulation of the Spatial Distribution of Hydraulic Conductivity in Porous Media through Different Methods, Math. Probl. Eng., 2017(Article ID 4321918), 1–10, doi:10.1155/2017/4321918, 2017.

Zhang, Y. and Schaap, M. G.: Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3), J. Hydrol., 547(January 2017), 39–53, doi:10.1016/j.jhydrol.2017.01.004, 2017.

Zhao, C., Jia, X., Nasir, M. and Zhang, C.: Catena Using pedotransfer functions to estimate soil hydraulic conductivity in the Loess Plateau of China, Catena, 143, 1–6, doi:10.1016/j.catena.2016.03.037, 2016.

Table 1. Available environmental covariates.

Name	Resolution	Description
Soil		
soil type	100 m	according to Hungarian classification system (Pásztor et al., 2018a)
clay, silt, sand content	100 m	0-30, 30-60, 60-90 cm (Laborczi et al., 2018)
organic matter content	100 m	0-30 cm (Szatmári and Pásztor, 2018)
calcium carbonate content	100 m	0-30 cm (Pásztor et al., 2018b)
pH in water	100 m	0-30 cm (Pásztor et al., 2017)
Parent material	1:100000	(Gyalog and Síkhegyi, 2005), map was converted to raster layer
Topography		
digital elevation model	25 m	(Bashfield and Keim, 2011) elevation, slope angle, aspect, northing and easting aspects, planar curvatures, profile curvatures, combined curvatures, topographic position indices, topographic position indices, terrain ruggedness indices, roughness, dissection, surface to area ratio, multi-resolution valley bottom flatness, multi-resolution ridge top flatness, negative openness, positive openness, convergence indices, LS factor, vector ruggedness measure, surface convexity, flow accumulation area, flow length, topographic wetness indices by single and multi-flow algorithms, vertical distance to existing water bodies, vertical distance to existing water bodies, horizontal distance to existing water bodies, smoothed version of elevation, smoothed version of profile curvature, smoothed version of slope, smoothed version of total curvature, standard deviations of elevation, standard deviations of profile curvature, standard deviations of slope, standard deviations of total curvature
Climate		
WorldClim	30”	(Fick and Hijmans, 2017) mean monthly temperature, precipitation, solar radiation, water vapour pressure, mean monthly minimum and maximum temperature
Hungarian data	100 m	(Szentimrey and Bihari, 2007) The spatial layers were compiled using the MISH method elaborated for the spatial interpolation of surface meteorological elements based on a 30 year observation ~~of~~ by the Hungarian Meteorological Service with 0.5’ resolution. mean annual precipitation and temperature
State of vegetation		
MODIS	250 m	(Vermote, 2015) normalized difference vegetation index, near infrared, red
Land cover		

Copernicus Pan-European High Resolution Layers	20 m	(CEC EEA, 2012) tree cover density, forest type, impermeable cover of soil, wetland, grassland
CORINE Land Cover	25 ha	(CEC EEA, 2012) natural grassland, land principally occupied by agriculture

Table 2. Description statistics of measured soil properties of the Balaton catchment

Soil propety	N	Minimum	Maximum	Mean	SD	Median
Clay content (100 g g^{-1})	1453	0.00	79.43	21.27	9.38	20.29
Silt content (100 g g^{-1})	1349	0.36	73.99	38.48	16.11	40.92
Sand content (100 g g^{-1})	1349	2.85	95.94	40.37	21.48	35.09
Organic matter content (100 g g^{-1})	1269	0.00	28.93	1.18	1.57	0.73
Calcium carbonate content (100 g g^{-1})	925	0.00	72.00	9.75	11.97	4.50
pH in water (-)	1445	3.61	9.38	7.14	0.98	7.29
Saturated water content (cm^3 cm^{-3})	1299	0.324	0.883	0.469	0.066	0.461
Water content at field capacity (cm^3 cm^{-3})	1294	0.032	0.640	0.314	0.083	0.320
Water content at wilting point (cm^3 cm^{-3})	1284	0.006	0.462	0.167	0.075	0.160

Table 3. Performance of hydraulic PTFs on training and test datasets. THS: saturated water content, FC: field capacity, WP: wilting point, RF: random forest method, GBM: generalized boosted regression method, TEST_CHEM set: test dataset in which chemical soil properties are available for the predictions, TEST set: test dataset, in which chemical soil properties are not necessarily available for the predictions, RMSE: root mean square error, R^2: determination coefficient.

Predicted soil hydraulic property		Selected method*	Train set**			TEST set			TEST_CHEM set		
			R^2	RMSE (cm^3 cm^{-3})	N	R^2	RMSE (cm^3 cm^{-3})	N	R^2	RMSE (cm^3 cm^{-3})	N
THS	topsoil	GBM	0.453	0.052	5709	-	-	-	0.484	0.042	2448
		RF	0.488	0.041	5709	-	-	-	0.487	0.042	2448
	subsoil	GBM	0.429	0.045	8428	0.418	0.045	3611	0.400	0.046	2448
		RF	0.480	0.043	8428	0.429	0.045	3611	0.408	0.045	2448
FC	topsoil	GBM	0.714	0.043	5635	-	-	-	0.770	0.039	2416
		RF	0.736	0.041	5635	-	-	-	0.766	0.039	2416
	subsoil	GBM	0.738	0.044	8352	0.739	0.042	3579	0.751	0.040	2416
		RF	0.756	0.042	8352	0.746	0.042	3579	0.759	0.040	2416
WP	topsoil	GBM	0.722	0.038	5736	-	-	-	0.739	0.037	2459
		RF	0.736	0.037	5736	-	-	-	0.762	0.035	2459
	subsoil	GBM	0.717	0.041	8425	0.716	0.039	3611	0.711	0.038	2459
		RF	0.747	0.039	8425	0.737	0.038	3611	0.744	0.036	2459

* Input parameters included in all analysis for topsoils: ~~Hungarian~~ soil type according to Hungarian classification system, sand (50–2000 µm), silt (2–50 µm) and clay content (<2 µm) (100 g g^{-1}), mean depth (cm) and information on topography, vegetation, meteorology and parent material listed in Table 1. For subsoils organic matter content (100 g g^{-1}); pH in water and calcium carbonate content (100 g g^{-1}) were included as well.
** Prediction error calculated on training is based on out of bag error in case of RF and 5-fold cross-validation in case of GBM method.

Table 4. Performance of random forest method and parameters of the fitted variogram models during the geostatistical mapping approach.

Predicted soil hydraulic properties	Depth (cm)	Random forest			Variogram			
		R^2	RMSE (cm^3 cm^{-3})	N	Partial sill	Type	Range	Nugget
THS	0-30	0.403	0.055	324	0	„Nug”	-	32.552
	30-60	0.251	0.055	321	11.037	„Exp”	1531	18.357
	60-90	0.189	0.060	315	14.150	„Exp”	8211	27.067
FC	0-30	0.562	0.053	324	0	„Nug”	-	29.895
	30-60	0.532	0.056	321	0	„Nug”	-	26.539
	60-90	0.478	0.063	315	0	„Nug”	-	32.356
WP	0-30	0.463	0.052	324	0	„Nug”	-	23.689
	30-60	0.474	0.051	321	0	„Nug”	-	22.655
	60-90	0.466	0.056	315	32.718	„Sph”	2149	0

Table 5. Performance of soil hydraulic maps derived by random forest and kriging method (RFK), Hungarian pedotransfer functions (HUN-PTF) and from EU-SoilHydroGrids 250m dataset (EU-SHG) on the Balaton catchment. RMSE: root mean square error, SS_{mse}: mean square error skill score.

Predicted soil hydraulic property	Depth	Method	N	RMSE (cm^3 cm^{-3})	SS_{mse}	Sign. difference*
THS	0-30 cm	RFK	324	0.056	0.382	b
		HUN-PTF	350	0.067	0.118	b
		EU-SHG	348	0.070	0.041	a
	30-60 cm	RFK	321	0.060	0.119	a
		HUN-PTF	345	0.058	0.150	b
		EU-SHG	343	0.063	-0.004	a
	60-90 cm	RFK	315	0.063	0.112	b
		HUN-PTF	337	0.060	0.171	c
		EU-SHG	335	0.071	-0.149	a
FC	0-30 cm	RFK	324	0.053	0.547	b
		HUN-PTF	350	0.067	0.265	b
		EU-SHG	348	0.076	0.070	a
	30-60 cm	RFK	321	0.057	0.515	b
		HUN-PTF	345	0.069	0.278	b
		EU-SHG	343	0.084	-0.069	a
	60-90 cm	RFK	315	0.062	0.485	b
		HUN-PTF	337	0.074	0.232	b
		EU-SHG	335	0.095	-0.243	a
WP	0-30 cm	RFK	324	0.052	0.453	b
		HUN-PTF	349	0.062	0.244	ab
		EU-SHG	347	0.071	-0.038	a
	30-60 cm	RFK	321	0.052	0.467	b
		HUN-PTF	344	0.065	0.152	b
		EU-SHG	342	0.074	-0.112	a
	60-90 cm	RFK	315	0.057	0.443	c
		HUN-PTF	335	0.067	0.208	b
		EU-SHG	333	0.076	-0.026	a

*Different letters indicate significant differences at 0.05 level between the accuracy of the methods based on squared error, e.g. performance indicated with letter c is significantly better than the one noted with letter b and a.

Table 6. Proportion of mapped area having smaller than 0.025, 0.025-0.050, 0.05-0.100 and bigger than 0.10 cm^3 cm^{-3} absolute difference between predicted soil hydraulic values derived by geostatistical method (RFK) and applying pedotransfer functions on local soil and environmental covariates (HUN-PTF).

Absolute difference between RFK and HUN-PTF (cm^3 cm^{-3})	**Depth (cm)**	**% of mapped area**		
		THS	**FC**	**WP**
0-0.025	0-30	76	80	71
	30-60	86	77	65
	60-90	75	72	71
0.025-0.050	0-30	21	17	25
	30-60	10	21	26
	60-90	21	22	24
0.050-0.100	0-30	3	3	4
	30-60	4	2	9
	60-90	4	6	5
0.100 <	0-30	0	0	0
	30-60	0	0	0
	60-90	0	0	0

Table 7. Differences between pedotransfer function based (PTF) and geostatistical (RFK) mapping methods based on calculating saturated water content, field capacity and wilting point for the Balaton catchment.

Aspects of mapping	**Differences between the soil hydraulic mapping methods**	
	PTF – indirect method	**RFK – direct method**
Main steps of mapping	1. derive PTFs on available soil hydraulic dataset or use an appropriate PTF available from the literature, 2. apply PTFs on available environmental covariates	1. harmonize soil profile dataset available for the mapping based on required soil depth, 2. predict deterministic component, 3. calculate the residuals, estimate their variograms, krige them, 4. add kriged residuals to the deterministic component
Dataset used to describe relationship between soil hydraulic data and covariates	- any soil hydraulic dataset which is hydropedologically similar to the area for which soil hydraulic maps are required - advantages: mapping can be applied even if no soil hydraulic data is available for the study area; available PTF ~~also~~ can also be used - disadvantages: a soil hydraulic dataset is needed which has to be similar to data of the study site from soil hydropedological point of view; or if PTF is already available the soil hydrological dataset used to train the PTF has to be similar to the study site	- soil hydraulic data available for the catchment - advantages: soil hydraulic data is characteristic for the study site, locally extreme values can be better characterized - disadvantages: density of measured soil hydraulic properties available for the study site might not satisfy the needs for mapping; further to the soil property, which is mapped, measured data of soil properties used in the prediction of the deterministic component (e.g. particle size distribution, ~~OM~~ organic matter content) is required as well
Inclusion of soil depth	- can be included as independent variable - advantages: measured soil hydraulic properties are related to measured soil properties; soil hydraulic properties at any depth can be calculated - disadvantages: certain depths can be underrepresented in the training dataset which might increase prediction uncertainty	- in 2D kriging soil data (chemical, physical, hydraulic) is first harmonized in training dataset by splining to derive data for fix depth - disadvantages: measured soil properties are splined therefore calculated soil hydraulic properties are related to calculated soil properties, thus map relationship between them is derived from interpolated (namely splined) values
Spatial inference	- this method relies on the interpolation included in the input layers used for the mapping, thus the mapping is indirect - advantage: no further geostatistical analysis is needed to provide 3D information	- directly the soil hydraulic properties are interpolated - advantage: uncertainty of input layers is decreased due to adding the kriged residuals to the predicted values

	- disadvantage: uncertainty of input layers increase uncertainty of predicted soil hydraulic properties	
Information on uncertainty	- interpreted as the uncertainty of the PTFs - advantage: can be easily computed for PTFs - disadvantages: not location specific, but depends on the input parameter combination, uncertainty of input layers has to be added to the uncertainty of PTFs to provide information on the uncertainty of soil hydraulic maps, uncertainty of input environmental covariates is hardly definable if e.g. 60-70 of them are used for the mapping	- can be derived with e.g. bootstrapping - advantages: location specific; the uncertainty accounts for both the unexplained stochastic variation and the uncertainty in estimating the deterministic model - disadvantages: computationally demanding; require massive storage capacity; uncertainty of input layers has to be added to the uncertainty of RFK

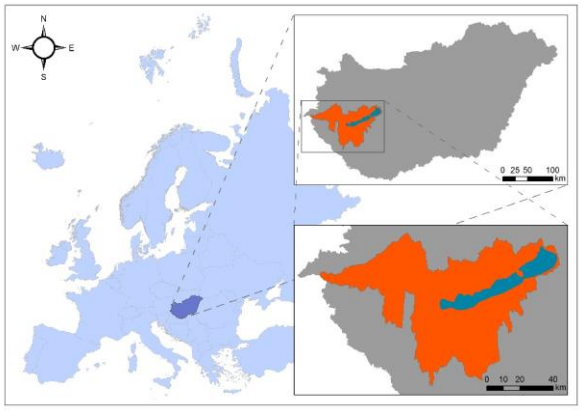


Figure 1. Location of study site Balaton catchment.

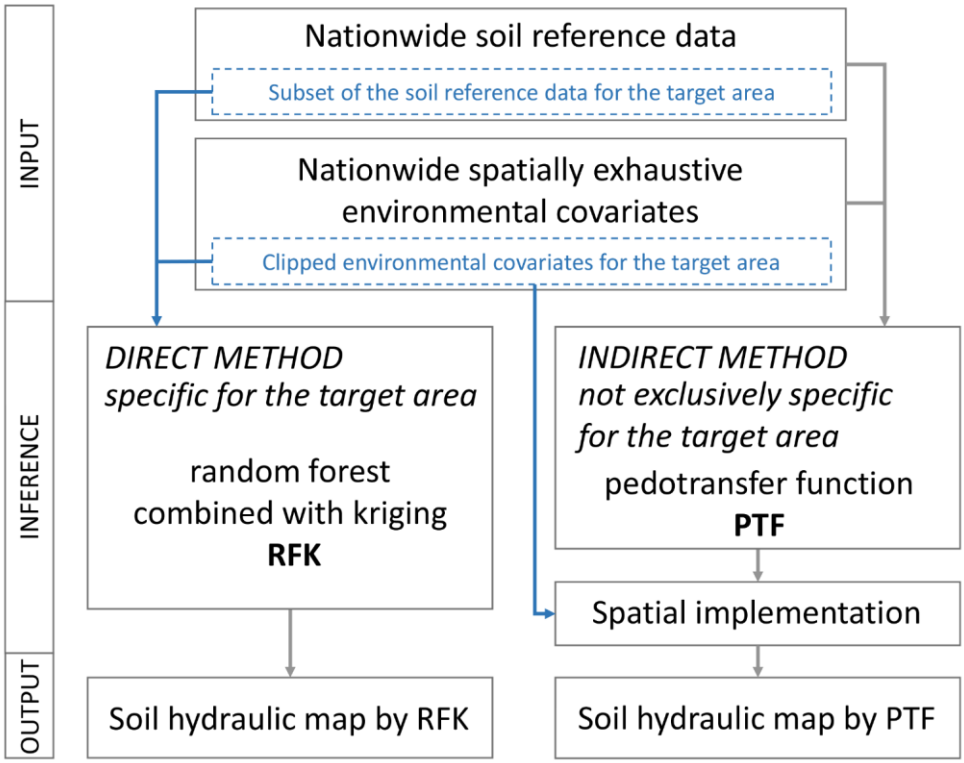


Figure 2. Flowchart about the main steps of direct and indirect soil hydraulic mapping methods.

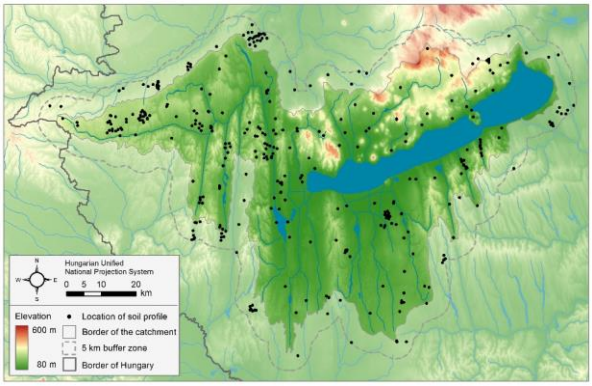


Figure 23. Location of soil profile used for the geostatistical soil hydraulic mapping on the Balaton catchment study area. Solid line indicates border of the catchment, dashed line shows area with the 5 km buffer zone.

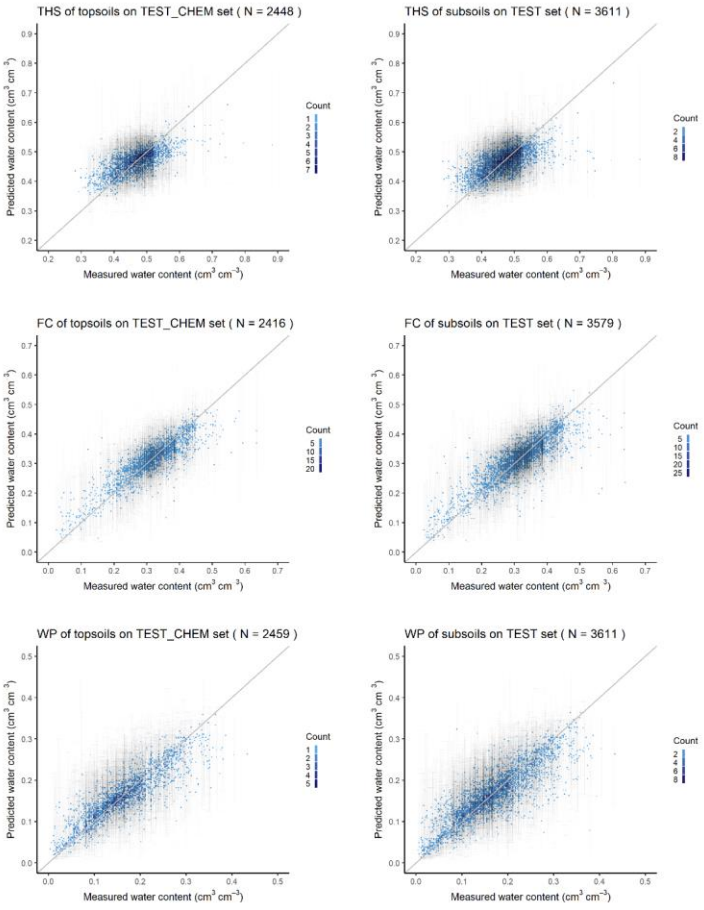


Figure 4. The scatter plot of the measured versus predicted water retention values with 90% prediction interval on test data sets based on random forest method. THS: saturated water content, FC: water content at field capacity, WP: water content at wilting point, TEST_CHEM set: test dataset in which chemical soil properties are available for the predictions, TEST set: test dataset, in which chemical soil properties are not necessarily available for the predictions.

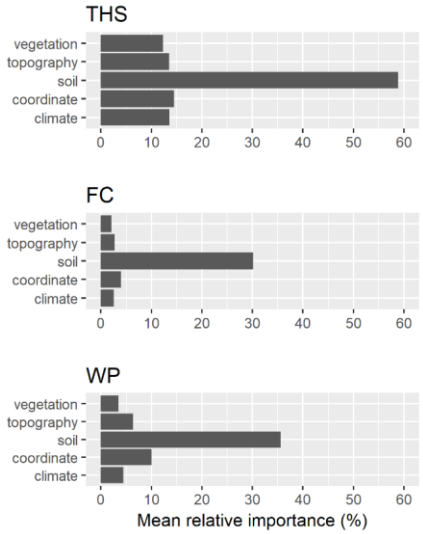


Figure ~~3~~5. Mean relative importance of covariates used to predict soil hydraulic properties based on random forest analysis on the training set of MARTHA database. THS: saturated water content, FC: water content at field capacity, WP: water content at wilting point.

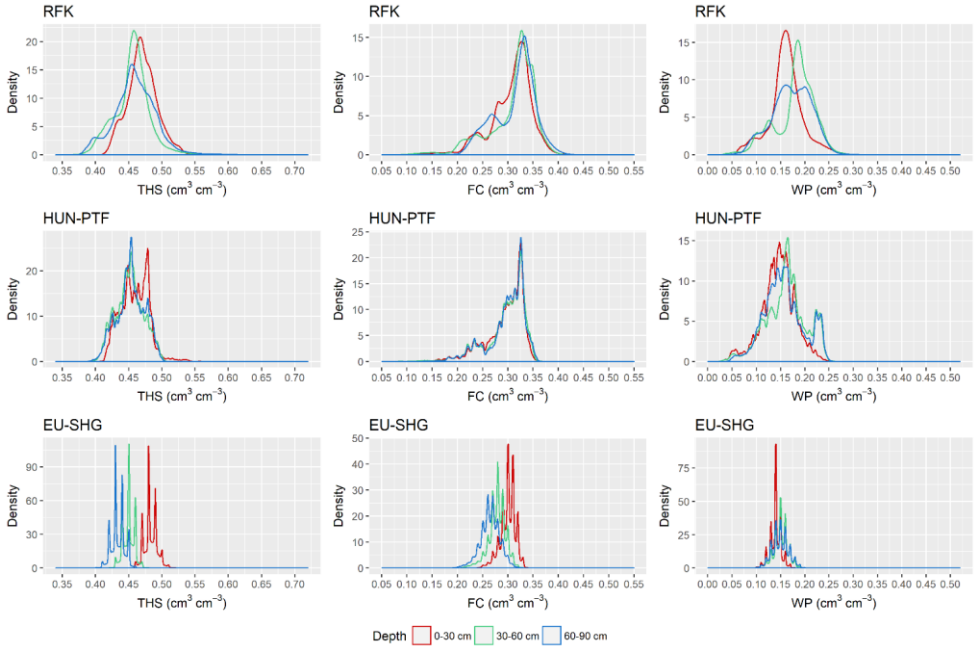


Figure 64. Density plots of mapped soil hydraulic values by mapping methods and depth. THS: saturated water content, FC: water content at field capacity, WP: water content at wilting point, RFK: derived by random forest with kriging, HUN-PTF: calculated with Hungarian pedotransfer functions, EU-SHG: values from EU-SoilHydroGrids 250m 3D European soil hydraulic mapsdataset.

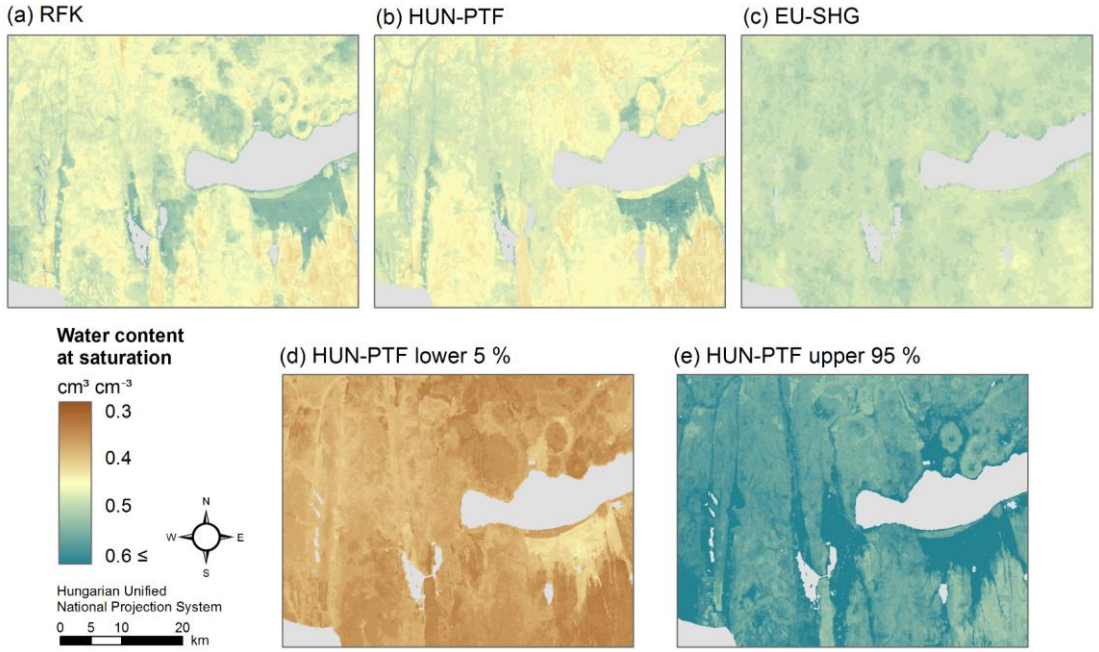


Figure 7. Map of water content at saturation in 0-30 cm soil depth derived by random forest and kriging mapping approach (RFK) (a), Hungarian pedotransfer functions (HUN-PTF) (b) and cut from the EU-SoilHydroGrids 250m dataset (EU-SHG) (c), possible lower 5 % (d) and upper 95 % (e) based on HUN-PTF for a section of the Balaton catchment.

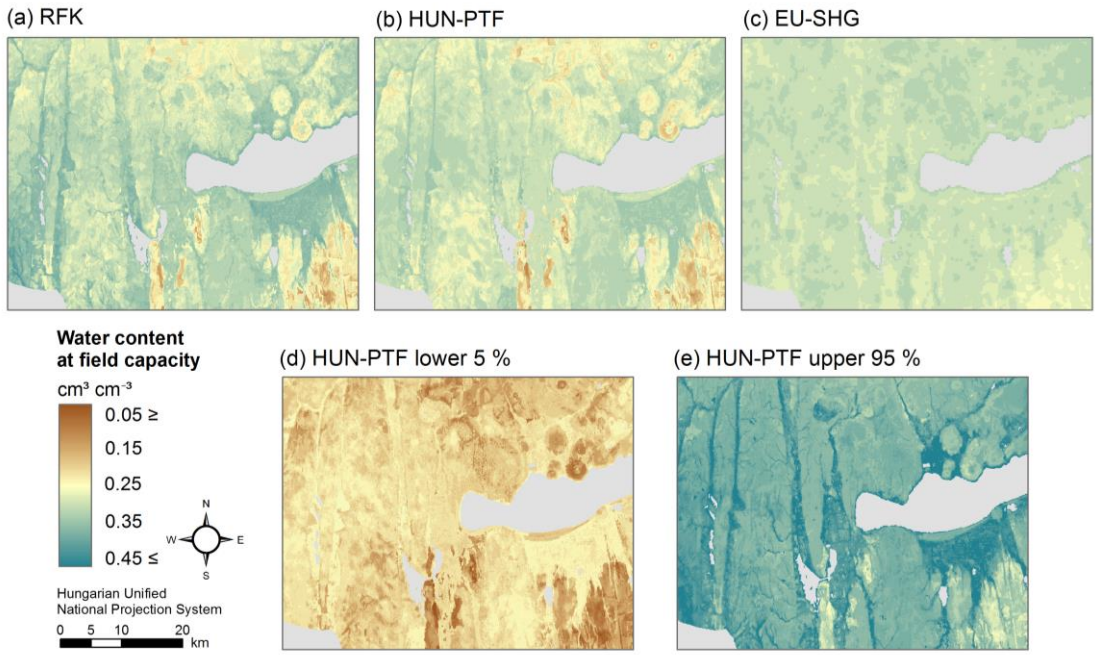
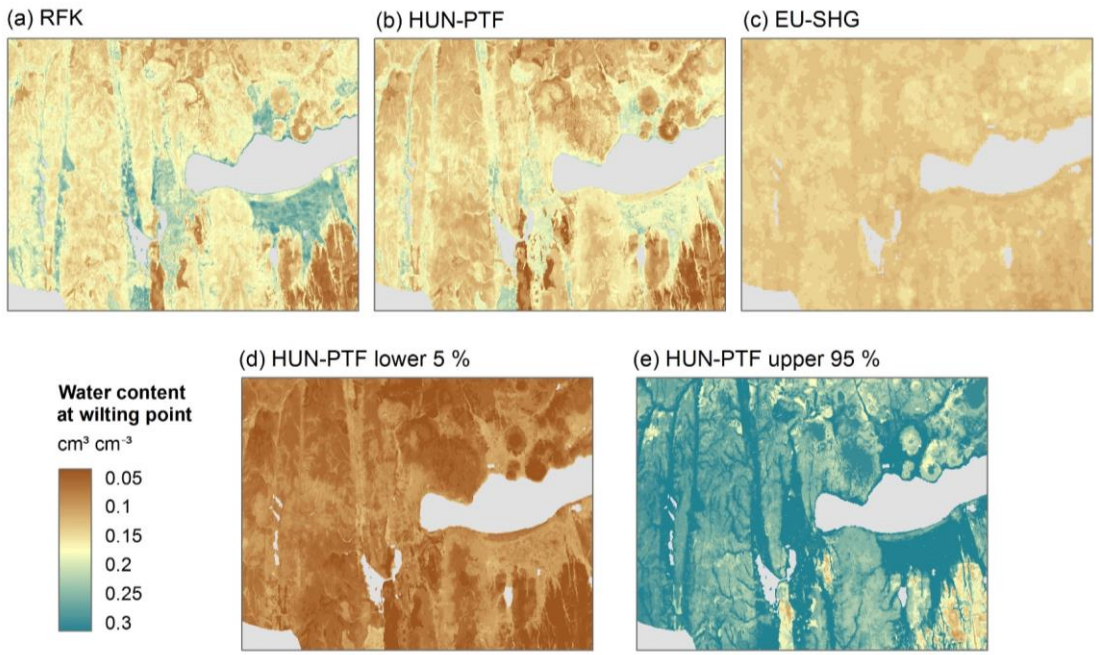


Figure 8. Map of water content at field capacity in 0-30 cm soil depth derived by random forest and kriging mapping approach (RFK) (a), Hungarian pedotransfer functions (HUN-PTF) (b) and cut from the EU-SoilHydroGrids 250m dataset (EU-SHG) (c), possible lower 5 % (d) and upper 95 % (e) based on HUN-PTF for a section of the Balaton catchment.


(a) RFK
(b) HUN-PTF
(c) EU-SHG
(d) HUN-PTF lower 5 %
(e) HUN-PTF upper 95 %
Water content at wilting point
cm³ cm⁻³
0.05
0.1
0.15
0.2
0.25
0.3

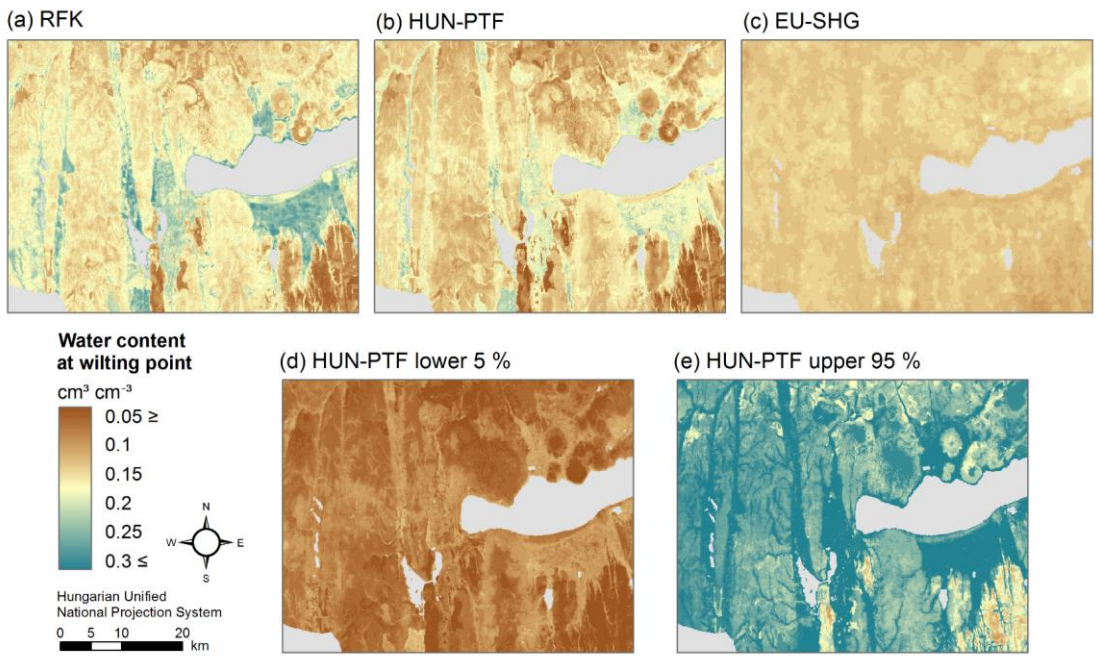
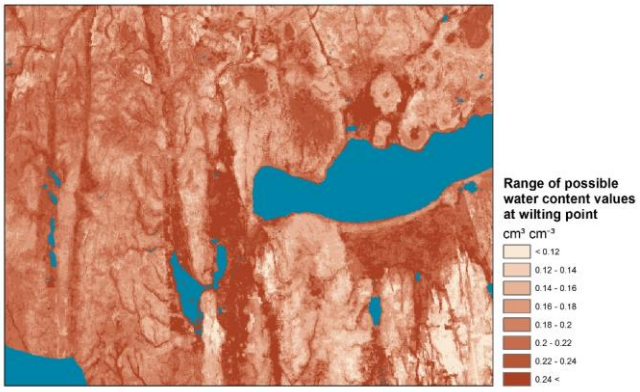


Figure 59. Map of water content at wilting point in 0-30 cm soil depth derived by random forest and kriging mapping approach (RFK) (a), Hungarian pedotransfer functions (HUN-PTF) (b) and cut from the EU-SoilHydroGrids 250m 3D European soil hydraulic dataset (EU-SHG) (c), possible lower 5 % (d) and upper 95 % (e) based on HUN-PTF for a section of the Balaton catchment.


Range of possible water content values at wilting point
cm³ cm⁻³
< 0.12
0.12 - 0.14
0.14 - 0.16
0.16 - 0.18
0.18 - 0.2
0.2 - 0.22
0.22 - 0.24
0.24 <

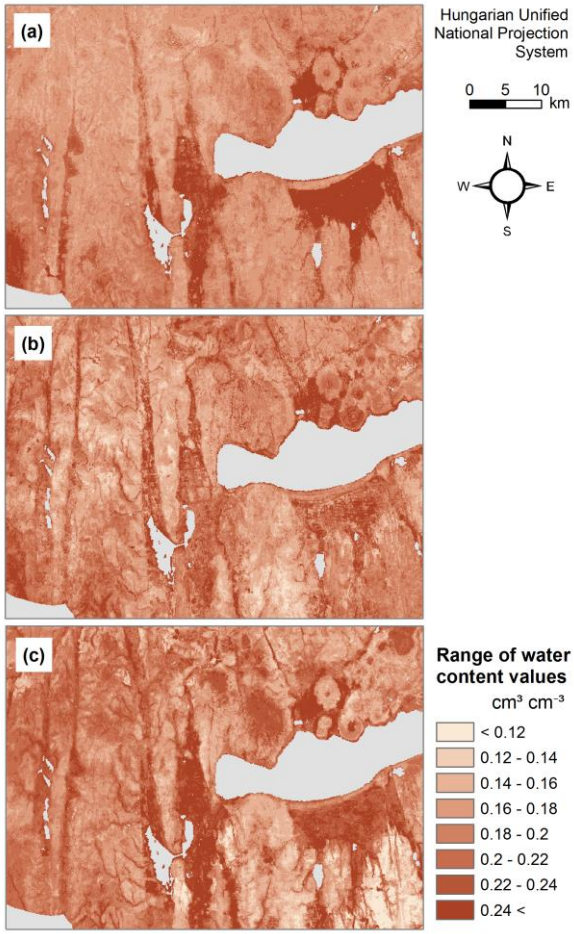


Figure 106. Differences between possible lower 5 % and upper 95 % water content at saturation (a), field capacity (b) and wilting point (c) in 0-30 cm soil depth for a section of the Balaton catchment.