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

Brigitta Tóth[1,2], Gábor Szatmári[1], Katalin Takács[1], Annamária Laborczi[1], András Makó[1], Kálmán
5 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)

10 **Abstract.** Spatial 3D information on soil hydraulic properties for areas larger than plot scale are usually derived with indirect
methods due to lacking measured information on those. Soil hydraulic properties are calculated with applying pedotransfer
functions (PTFs) – which describe the relationship between the desired soil hydraulic parameter and easily available soil
properties determined on a soil hydraulic point dataset – on available soil maps. Our aim was to analyse difference in
performance and spatial patterns between soil hydraulic maps derived with indirect (using PTFs) and direct (geostatistical)
15 mapping methods. We performed the study on Balaton catchment in Hungary, where density of measured soil hydraulic data
fulfils the requirements of geostatistical methods. Maps of saturated water content (THS), field capacity (FC) and wilting point
(WP) for 0-30, 30-60 and 60-90 cm soil depth were prepared. PTFs were derived with random forest method on the whole
Hungarian soil hydraulic dataset (MARTHA: soil chemical, physical, taxonomical and hydraulic information of some 12000
samples) complemented with information on topography, climate, parent material, vegetation and land use. As a direct method
20 random forest combined with kriging (RFK) was applied on 359 MARTHA soil profiles located in the Balaton catchment.
There was no significant difference 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 at 30-60 and
60-90 cm soil depth, in case of WP the RFK outperformed the PTFs at 60-90 cm depth. Difference between the PTF based
25 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 mapping
of the soil hydraulic properties is less computational intensive. Detailed comparison of the maps derived by the PTF based and
the RFK is presented in the 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 neither at regional or continental scale. Analyses on the prediction of soil hydraulic properties has been started

5 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 performance of predictions (pedotransfer functions - PTFs) when newer statistical methods and/or new data become available. Latest works on it among others are McNeill et al. (2018); Román Dobarco et al. (2019); Zhang and Schaap (2017).

10 Tree based machine learning algorithms (MLA) have been found 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 ensembles of trees, providing prediction of several individual trees built with 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 extraordinary values, therefore predictions will be always reasonable if

15 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 a single MLA, which consist of several models through bagging or boosting of e.g. decision tree or support vector machine or neural network algorithm, but can consist of different models and derive the average of all. It has been shown, that the more models are combined for the prediction the more accurate the result is (Baker

20 and Ellison, 2008; Cichota et al., 2013; Nussbaum et al., 2018; Wu et al., 2018). Hengl et al. (2017) also used merged ensemble predictions by calculating the weighted average of two MLAs 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.

25 Tree type ensemble algorithms were found to be successful for 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 not yet intensively applied to derive input parameters for hydrological modelling (Koestel and Jorda, 2014; Tóth et al., 2014).

Hengl et al. (2018a) tested several MLAs to map potential natural vegetation. From those random forest performed the best.

30 Nussbaum et al. (2018) analysed different methods to map several soil properties for three study sites in Switzerland. They also found that random forest method performed the best when single model is used. Results of Rudiyanto et al. (2018) also showed that among several tested MLAs tree-based models performed the best. Hengl et al. (2018b) reviewed MLAs and

geostatistical methods for soil mapping and found that 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 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 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 field capacity and wilting point with geostatsitical methods for an area of 1483 ha. They optimized number of measurements 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. 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 uncertainty of the prediction of soil hydraulic properties.

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 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 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 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

5 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 for 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 parts. 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,

10 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, soil chemical properties – OM,

15 CaCO$_3$ and pH – were available only for the 0-30 cm depth, therefore those could be considered only for the topsoil predictions. Information on topography, meteorology, geology and vegetation listed in Table 1 was used as environmental covariates for the elaboration of PTFs as well as for mapping.

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 Hungarian Uniform National

20 Projection system, rasterized if necessary and resampled to 100 m resolution.

2.2 Soil hydraulic dataset

For the analysis of the relationship between soil hydraulic properties and other environmental covariates (173) we used the Hungarian Detailed Soil Hydrophysical Database (Makó et al., 2010) extended with topographical, meteorological, geological information and remotely sensed vegetation properties (Table 1), called MARTHA ver 3.0 (acronym of the Hungarian name

25 of the dataset). MARTHA consists of 15142 soil horizons belonging to 3970 soil profiles.

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

30 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

5   respectively.

2.3 Mapped soil hydraulic properties

We mapped the most often used soil water retention values, water content at 0, -330 and -15000 cm matric potential values, THS, FC and WP respectively. Definition of FC varies across different countries. In Hungary FC is determined at -300 cm matric potential, therefore water content at -100 or -200 cm was not analysed in the presented work.

10 The information on soil properties were available for 0-30, 30-60 and 60-90 cm soil depths 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 mapping

2.4.1 Pedotransfer function based indirect mapping (HUN-PTF)

15 We analysed prediction performance of the two most efficient 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 the possibility to calculate quantiles during the prediction, in this way prediction uncertainty can be provided based on parameter input combination.

Both algorithms build ensemble of models from regression trees. The difference is the way of building the forest from the

20 individual trees. RF relies on averaging the result of the trees in the ensemble. The trees are grown independently from each other (Breiman, 2001), therefore it is a bagging type ensemble. 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).

Optimization of parameter set in RF and GBM model was performed with the train function of R package 'caret' (Kuhn et al.,

25 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 ('mtry') was tuned. In case of GBM influence of interaction depth and shrinkage were analysed. In ranger RF default value is 500 for number of trees, that was used for both RF and GBM. Also for minimum number of observations in 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 during the tuning of model parameters.

Most important 50 independent variables have been selected from both GBM and RF models, then parameter tuning was performed again with the decreased number of input variables. We compared the accuracy of all models based on the cross-validation results and built the final prediction model (PTF) with the best performing and simplest algorithm on all training data with the optimized parameters. Performance of the PTFs was described with root mean square error (RMSE) (Eq. 1) and

5   coefficient of determination (R^2) (Eq. 2).

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

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

Performance of PTFs on the training dataset was assumed based on the results of five-fold cross-validation, and out-of-bag samples for GBM and RF respectively. In RF accuracy on out-of-bag was analysed. Uncertainty of the predictions was

10   characterized with the 5 and 95% quantiles of the predicted values, calculated within the 'ranger' and 'gbm' packages during deriving 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), hence those were mapped

15   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 pixels based, further to the median value 5 and 95% quantiles of the predicted values were mapped for each soil hydraulic property based on the PTFs.

20   **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 RF introduced above, whereas the stochastic part of variation is modelled by kriging using the derived residuals.

For the geostatsitical analysis those samples of the MARTHA database were selected which fall within the catchment plus 5

25   km buffer zone area. The buffer zone was used to increase 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. 2). Table 2 shows 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-

30   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 – listed in Table 1 – have been selected and the final RF model was optimized with those covariates. 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

5 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 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

10 Performance of soil hydraulic maps was evaluated based on measured soil hydraulic properties calculated for 0-30, 30-60 and 60-90 cm depth with 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 maps.

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

Performance of soil hydraulic maps derived with HUN-PTFs and RFK was compared to the 3D European soil hydraulic maps

15 (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-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.

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

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

25 ## 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 difference between RMSE values were less than 0.0001 cm^3 cm^{-3}, which is negligible from a practical point of view. In Nussbaum et al. (2018) number of input parameters were decreased from 300-500 environmental covariates to 10, 20, 30, 40, 50 most important ones. 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

5   used for the predictions. PTFs derived with RF are not sensible for reducing independent variables to the most important ones, neither multicollinearities between independent variables decrease performance. 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 case of RF optimal number of input parameters randomly selected at each split was between 10 and 20, depending on soil

10   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, influence of shrinkage and increasing number of trees to 1000 was 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 1000 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

15   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 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

20   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 0.040 and 0.043 cm^3 cm^{-3} for THS, 0.040 and 0.042 cm^3 cm^{-3} for FC, 0.036 and 0.039 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 0.400-0.487, 0.739-0.770 and 0.711-762 for THS, FC and WP respectively on test

25   sets. When we compared performance of RF derived for topsoils – which includes OM, pH and CaCO$_3$ 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 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

30   et al., 2014; Hengl et al., 2017; Nussbaum et al., 2018). When other environmental covariates than soil are not included among input parameters chemical properties significantly improve prediction (Hodnett and Tomasella, 2002; Khodaverdiloo et al., 2011; Tóth et al., 2015). In case of THS range of predicted values using chemical parameters as well were closer to range of measured values, therefore we 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 OM, pH and CaCO₃ – 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 OM, silt, sand content, pH, clay, CaCO₃ content are the most important variables with relative importance 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 PSD and OM are the most important variables, having relative importance around and over 20 %. ST, 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 less than 5%. For WP longitude, mean monthly precipitation of November and July, elevation, vertical and horizontal distance to existing water bodies, CaCO₃, mean monthly radiation, pH, depth, mean monthly water vapour pressure, multi-resolution ridge top flatness and spectral reflectance in near infrared have relative importance 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 (OM, CaCO₃, pH) are not included among input parameters, sand, silt, clay content are far the most important three independent variables (39-100 %). In case of THS also depth has higher relative importance (52 %). For the prediction of FC importance of ST increases to 18 %. In case of WP 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. 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 number of environmental covariates – from 173 to 50 and from 170 to 50 in case of topsoil and subsoil, respectively – significantly improved the prediction accuracy for each soil hydraulic property. For the final RF models the optimal mtry 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, OM (for topsoil), clay, silt and sand content and the pH (for topsoil). The final RF models were used for estimating

5   the deterministic component for each soil hydraulic property.

The parameters of the fitted variogram models are summarized in Table 4. In 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

10   Lagacherie, 2017). 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.

15   **3.3 Performance of soil hydraulic maps**

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 several environmental covariates were considered for the predictions and relationship between easily available soil properties and soil hydraulic parameters were derived on local data.

20   In case of mapping six out of nine soil hydraulic maps there was no significant difference between maps derived by RFK and HUN-PTFs. In 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.

Range of predicted values is smaller in case of HUN-PTF method than in RFK, which is due to the "averaging approach" of the algorithm which in case of RFK is spatially corrected allowing wider range in the predicted values (Fig. 4, 5). Density plot

25   of predicted values are smoother in case of RFK than in HUN-PTF and EU-SHG maps (Fig. 4). This is due to adding residuals of kriging which modifies the values derived by random forest. In EU-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 maps. In SoilGrids algorithms are derived on a global dataset (Hengl et al., 2017), which has sparser measured data than the Hungarian soil profile database used to map soil properties

30   (Laborczi et al., 2018; Szatmári and Pásztor, 2018). Further to it RF is based on an averaging algorithm, these limits the ability to describe local extreme values. These result smaller range and variability of calculated soil hydraulic properties on EU-SHG maps than on RFK or HUN-PTF ones (Fig. 4) The basic Hungarian soil maps were derived with regression kriging methods, thus provide smoother soil input data for the calculations. As an example how differences in the range of predicted soil hydraulic properties can be visualized the map of WP is shown on Fig. 5 (a), (b), (c) for a selected area of the catchment.

Difference between the new and already available maps comes also from the difference in resolution, which is 100 m for RFK and HUN-PTF and 250 m for EU-SHG. Even though 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

5   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. Map of possible lower 5 % and upper 95 % values based on the HUN-PTF method are also shown in Fig. 5 (d), (e). The range between the lower and upper possible values (Fig. 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 PTF approach and MLA combined with geostatistics to derive soil hydraulic maps and compare their accuracy to that of EU-SoilHydroGrids. Therefore we didn't differentiated uncertainty of the maps originating from the soil input layers – i.e. DOSoReMI.hu and SoilGrids.

15   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

20   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 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.

25   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 provide broader

30   information and possibility for wider application of the PTFs. The presented HUN-PTF mapping method can be applied in any catchments of Hungary.

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
5 covariates available for the catchment. Then we compared their performance to the 250 m resolution 3D European soil hydraulic maps (EU-SoilHydroGrids).

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
10 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.

15 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,
20 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
25 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
30 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).

5 **Data availability.** The 3D soil hydraulic maps of Balaton catchment are freely available for non-commercial use from the Institute for Soil Sciences and Agricultural Chemistry Centre for Agricultural Research Hungarian Academy of Sciences (http://mta-taki.hu/en/kh124765/maps) 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
10 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.

15 **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

20 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,
25 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.

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.

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

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.

10 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

15 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,

20 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

25 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.

30 Learn., 40, 139–157, doi:10.1023/A:1007607513941, 2000.

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.

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.

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.

5 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.

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.,

10 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.

15 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.

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

20 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.

25 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.

30 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.

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, (November 2017), 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
5 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 30 year observation of 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, 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 (cm³ cm⁻³)	N	R^2	RMSE (cm³ cm⁻³)	N	R^2	RMSE (cm³ 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, 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 radom forest and kriging method (RFK), Hungarian pedotransfer functions (HUN-PTF) and from EU-SoilHydroGrids 250m (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 difference at 0.05 level between the accuracy of the methods based on squared error, e.g.

5   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 - advantage: mapping can be applied even if no soil hydraulic data is available for the study area; available PTF also can be used - disadvantage: 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 - advantage: soil hydraulic data is characteristic for the study site, locally extreme values can be better characterized - disadvantage: 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) is required as well
Inclusion of soil depth	- can be included as independent variable - advantage: measured soil hydraulic properties are related to measured soil properties; soil hydraulic properties at any depth can be calculated - disadvantage: 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 - disadvantage: 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 - disadvantage: uncertainty of input layers increase uncertainty of predicted soil hydraulic properties	- directly the soil hydraulic properties are interpolated - advantage: uncertainty of input layers is decreased due adding the kriged residuals to the predicted values
Information on uncertainty	- interpreted as the uncertainty of the PTFs - advantage: can be easily computed for PTFs	- can be derived with e.g. bootstrapping - advantage: location specific; the uncertainty accounts for both the unexplained stochastic

- disadvantage: 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

variation and the uncertainty in estimating the deterministic model

- disadvantage: computationally demanding; require massive storage capacity; uncertainty of input layers has to be added to the uncertainty of RFK
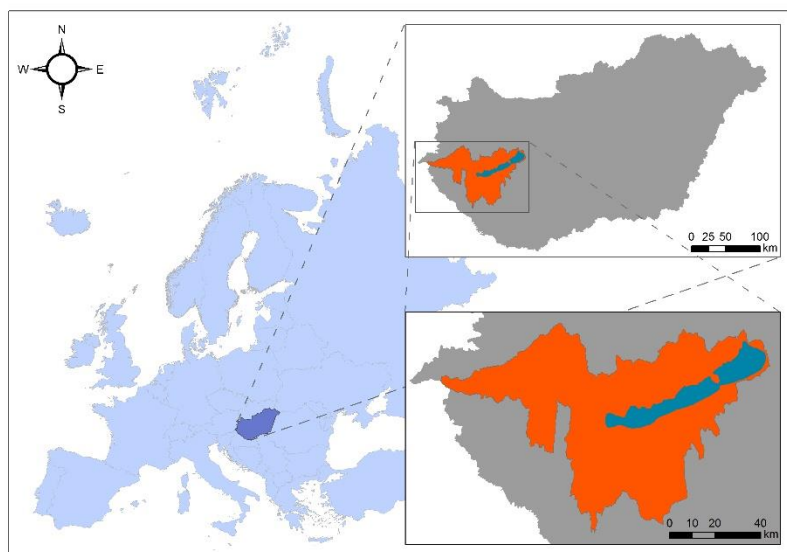
Figure 1. Location of study site Balaton catchment.
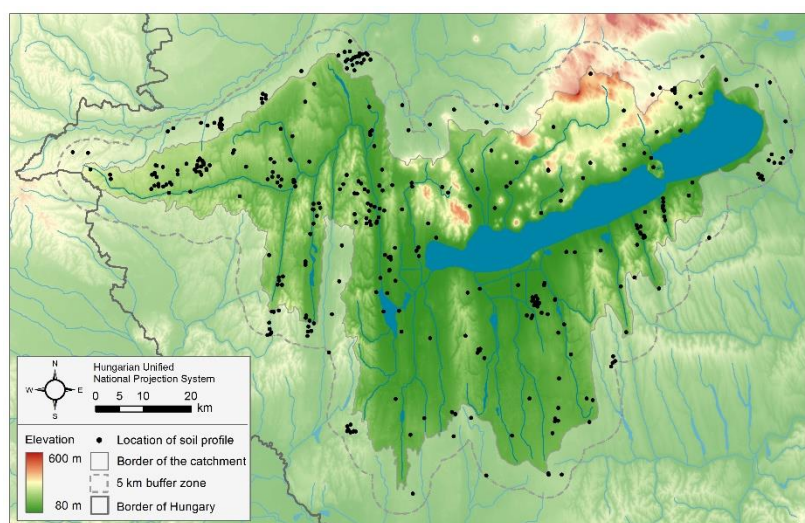
Figure 2. Location of soil profile used for the geostatistical soil hydraulic mapping on the Balaton catchment study area. Solid line indicates border of the catchment, dashed line shows area with the 5 km buffer zone.
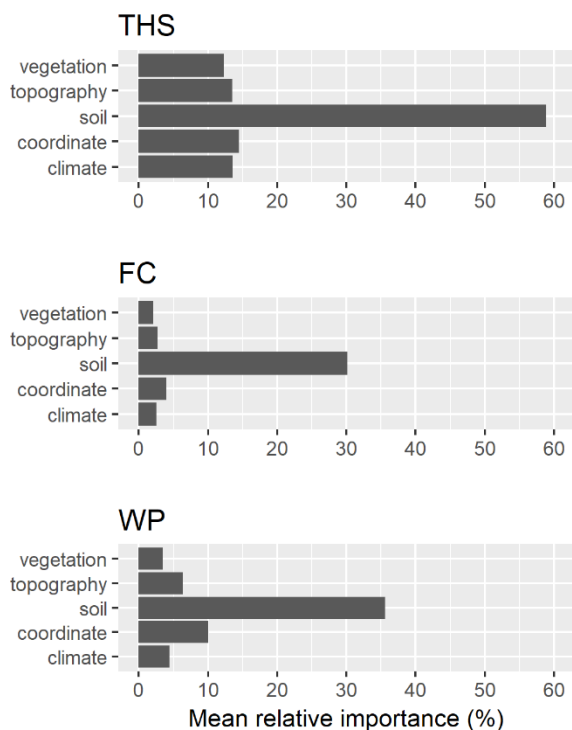
Figure 3. Mean relative importance of covariates used to predict soil hydraulic properties based on random forest analysis on the training set of MARTHA database. THS: saturated water content, FC: water content at field capacity, WP: water content at wilting point.
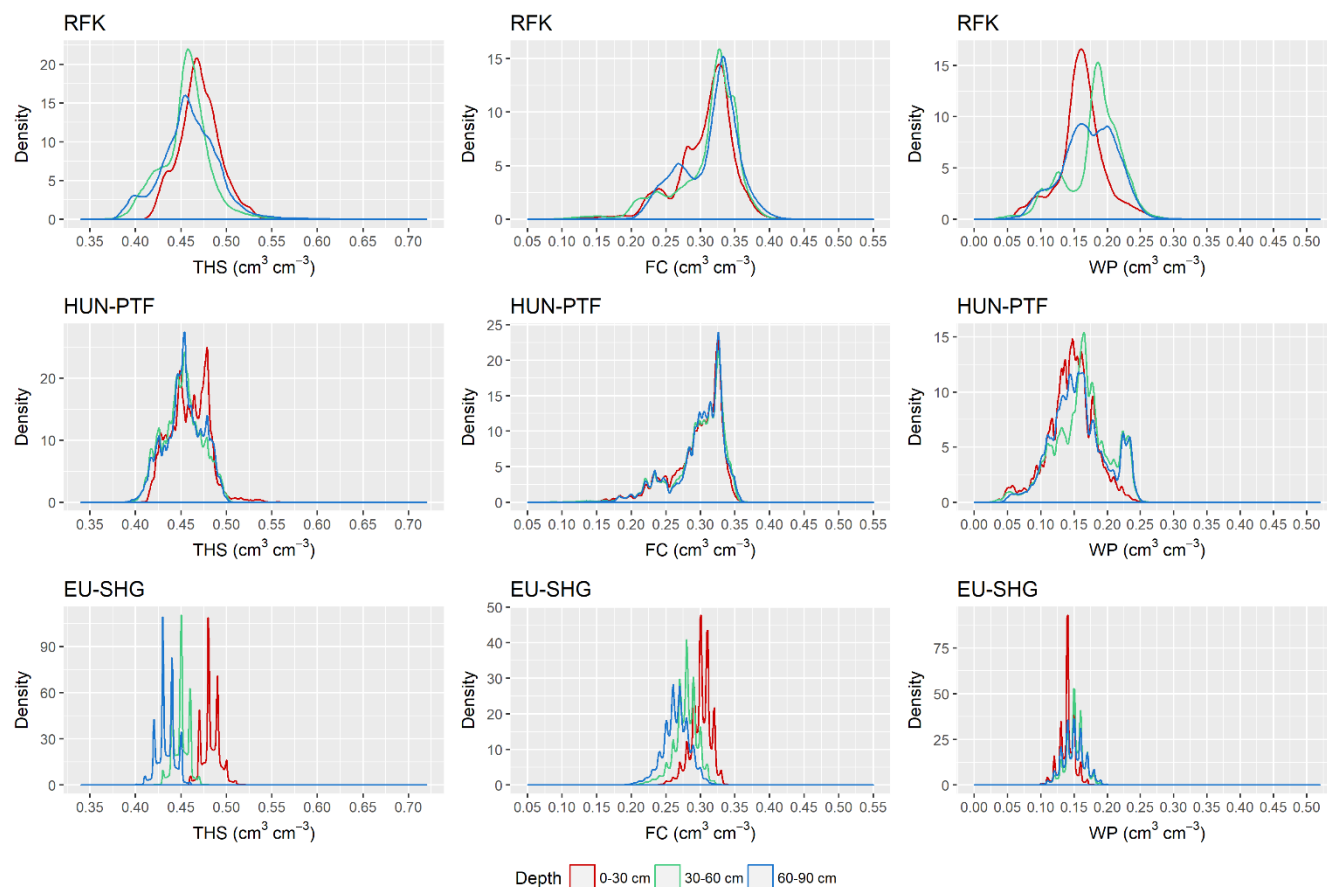
Figure 4. Density plots of mapped soil hydraulic values by mapping methods and depth. THS: saturated water content, FC: water content at field capacity, WP: water content at wilting point, RFK: derived by random forest with kriging, HUN-PTF: calculated with Hungarian pedotransfer functions, EU-SHG: values from 3D European soil hydraulic maps.
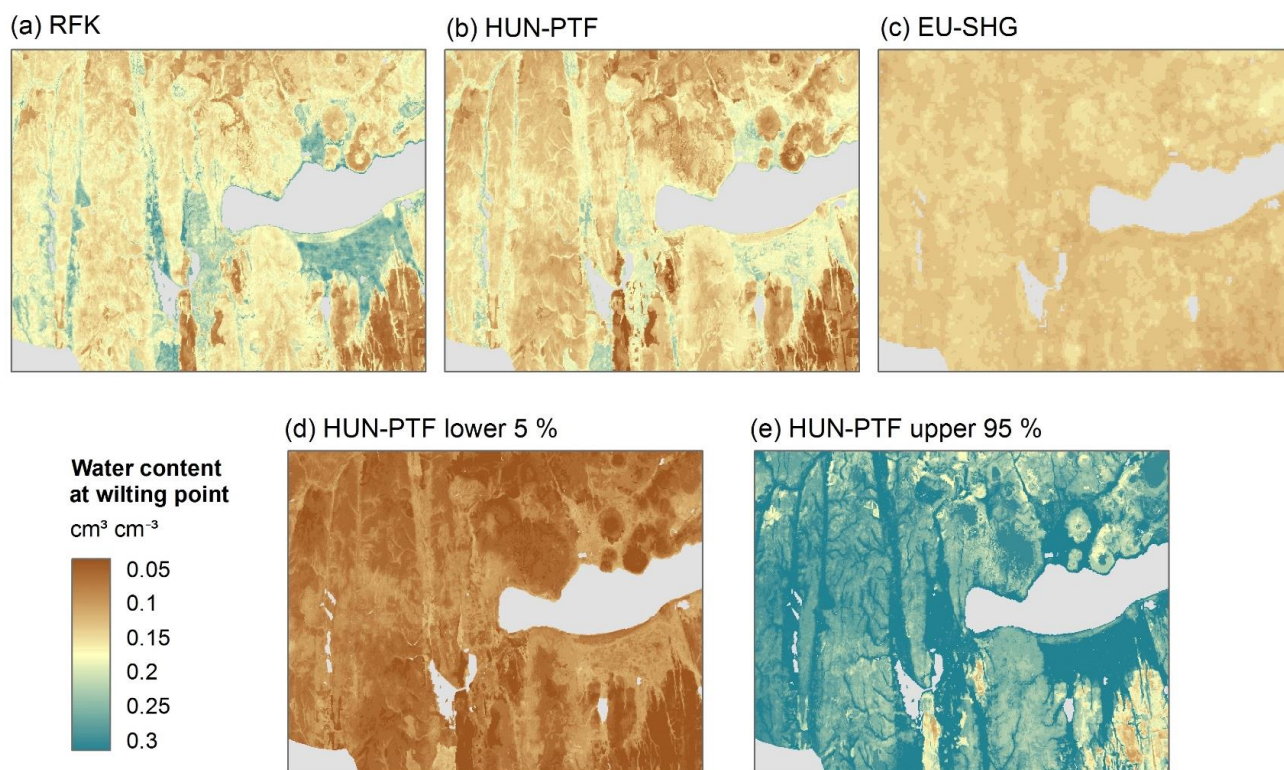
Figure 5. Map of water content at wilting point in 0-30 cm soil depth derived by random forest and kriging mapping approach (RFK) (a), Hungarian pedotransfer functions (HUN-PTF) (b) and cut from the 3D European soil hydraulic dataset (EU-SHG) (c), possible lower 5 % (d) and upper 95 % (e) based on HUN-PTF for a section of the Balaton catchment.
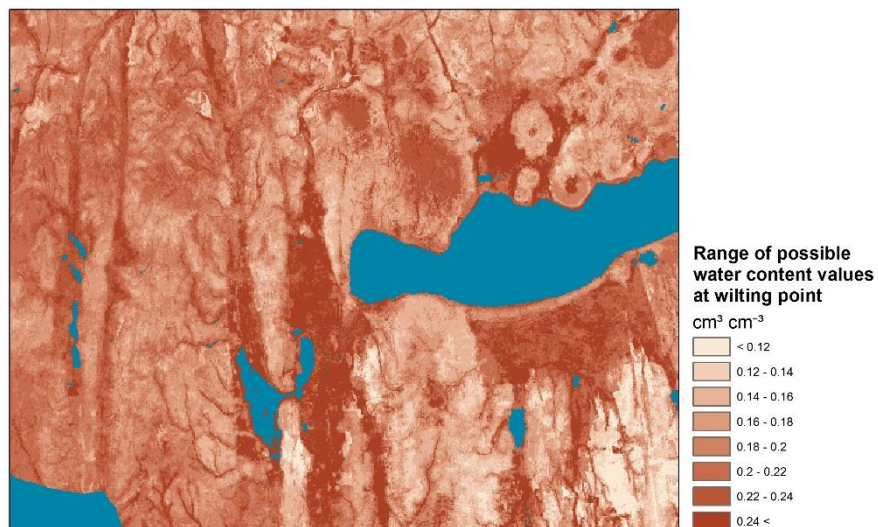
Figure 6. Difference between possible lower 5 % and upper 95 % water content at wilting point in 0-30 cm soil depth for a section of the Balaton catchment.