Topical Editor Decision: Reconsider after major revisions (24 Aug 2020) by Conrad Jackisch
Comments to the Author:
Dear Alexander and co-workers,

Thank you very much for your intense work on the manuscript. It really gained a lot more conciseness. I am now in the unpleasant situation to ask for another (hopefully largely technical) iteration before I can send it to the reviewers. The reason for this is that I anticipate rather superficial comments since the flow of the manuscript is not yet to the point. I am fully confident that the following suggestions are easy to work through and that it will help the next iteration to finally get the manuscript into a publishable shape. I apologise to cause you work but I am sure it is to the best of your manuscript.

We appreciate your input and your suggestions have definitely improved the manuscript a lot!

Abstract:
P1L22-24 (I refer the pages and line numbers to the annotated version in your reply letter) is the most important point here (in my view). The first paragraph appears to be more a relict of the first version. I also do not understand in which way the scales are of importance. Later in the manuscript you relate to SoilGrids as reference but this comparison is not at all summarised in the abstract. Moreover, I was wondering if the title might be amended to somehow relate to the large sources in addition to the "Soviet Soil Map".
What I have in mind is especially that the situation you worked through is not uncommon. There are soil maps in many countries but the eco-hydrological process properties/parameters are very difficult to distill. You provide one example how to do this plus the final data to use SWAT in Estonia. So the central aspect of your manuscript should really emphasise on this as an example.

Thank you for this recommendation. We took this into consideration and re-organized the manuscript.

Links as citations:
Although I can understand the notion of using the weblinks in the manuscript, I think the proper way would be to cite the sources similar to paper references. Websites will also require a date for the last access. Please see the guidelines here:
https://publications.copernicus.org/for_authors/manuscript_preparation.html

We reduced the web links in the ms and formatted them according to the journal guidelines.

Introduction:
P2L4-6 are central to me. I consider it a better start of the manuscript to state the problem instead of referring to SWAT. This also immediately opens your story: There is detailed soil information but eco-hydrological and pedi-hydrological process modelling (and understanding) requires information on different properties at finer scales. Then you can generalise what information exists in soil maps and global datasets and at what resolution.

We re-organized and re-wrote the introduction to better address the general problem and introduce the reader to the topic.

I prefer to give the scale triplet (Western and Blöschl, 1999) instead of a map scale which I have little

idea about the actual data density and validity. At the moment there are 1:10000, 30 arcsec, 250m, 1:500000, 1km etc. given. But I find it very difficult to get an idea about these resources, how the data has been derived and for which scale it might be applicable. I.e. remote sensing of soil properties in temperate climate is not very reliable nor can it give information about the soil profiles over depth. However they provide the advantage of spatially continuous data. You also point to the many soil sampling campaigns leading to the existing soil map. The soil samples are exactly opposite to remote sensing as they provide very detailed information for some points but it remains challenging to transfer this information to a spatially continuous map. I see this in close relation to the overview about data you use as additional inputs and data which is used as references plus the actual challenge of your study. Moreover, this is a fundamental link to our special issue I really think you should exploit for your argumentation.

Unfortunately, we cannot represent these datasets in scale triplet because all of them are representing continuous data and achieved in either generalisation, interpolation or statistical modelling. The final map scale is chosen based on the input data density (e.g. number of soil profiles per square km) and spatial resolution of environmental covariates that have been used for statistical modelling.

The introduction features so many possible data sources but it actually gives no outline of your manuscript. This comes P6L10ff. Somehow also 2.1 is hardly discernible from the introduction this way. I propose to restructure here. The introduction needs to guide the reader into your topic. The long list of soil-related resources can be the first subsection in the methods section. You may consider the many possible resources to be reported as a table? Please also clarify what information will be used to derive your dataset and which is simply reported for referencing and to clarify the importance of your study. Also the PTFs are not clear why they are given extensively in the introduction. I suspect that this can become one subsection in the methods and again detail the ones you really used and give the other references as brief as necessary.

We removed most of the dataset descriptions and web links from the introduction and rather focused on giving broader context and limitations of current fine resolution soil mapping.

Methods:
As stated above, Figure 1 should come rather early. As you have done, the methods then follow the four steps. And as proposed earlier, I suggest to include the respective background information, references etc. in these subsections. So everything about the actual mapping, unit derivation, etc. is maybe 2.1, 2.2 is A etc.

All Figures should be included as a vector graphics if possible. Fig. 1 still holds v1.0 and other details which might be easy to clarify in a revision. You said USLE_k is omitted as intermediate parameter? I would also suggest to reduce the used symbols to a minimum and to give a legend. The 4 main boxes could simply get their respective headers. The dashed sub-boxes and speech bubbles might simply become the same shape as annotations? The resulting properties could be given as something like a bottom line?
We simplified the figure and kept only the main workflow.

In the methods, the subsections should really align with the workflow in Fig. 1 to avoid confusion. If necessary, details can be of cause extended in subsubsections. Make it easy for the readers to follow. Always trace what information is really used and what is further reference. Please try to give

it always the same pattern: What is needed, available sources, other examples, how did you derive the property and how will you evaluate the results. Please avoid jumps between sub-topics. I think it is all there but still a little hard to fit the puzzle. E.g. 2.2 is about texture. Texture can be encoded as fractions of sand, silt and clay (or finer textural classes) or as WRB name. The layers (2.2.3) are for me a different thing more related to the geometric questions of mapping.

2.3 starts off with reference to 2.4. Maybe this could be more content of the overall introduction when Fig. 1 is explained in detail? Then 2.3 is clear to derive topographic indices and you have room to also link to the method background why these indices are selected. Maybe it is not even necessary to use so many subsections?

We removed the subsubsections and moved a lot of technical details to the supplements in order to keep the flow more straightforward for the reader.

In 2.4 you again motivate the application of PTFs of 2.5 but the overall pathway could be given earlier to avoid confusion and to really concentrate on SOC, SOM and BD. Since you use a random forest, I would simply remain with RF as one way of machine learning. Avoiding to use the two terms as synonyms could give room for details and clarity. I think that it is important to also clarify how and when PTFs (which are often also ML e.g. Rosetta as ANN) are used. Are they really part of 2.4?
2.5 is very brief although I suspect it to be the very crucial step. Contrary to the former subsections which should be more concise, I think here could be a little more explanation e.g. why the AWC, FC, WP are important. Moreover, again some layers are defined. This time at fixed depths. I still struggle to understand the geometric idea of your database.

We shortened the introduction and really try to focus only on PTFs and ML application used as PTF (such as in SoilGrids and EU-HydroSoilGrids) and only for our specific application domain.

I think the final database including the respective names and encodings could be a further section after the methods. There can come all details about your naming conventions etc.

Results:
I understand the results to align best along the four subsections again. At the moment I really struggle to follow how the evaluation is done and if the maps are somehow reliable. Maybe you could use a few of the sampling areas from Fig. 2 plus a few other areas with information from other resources here? The maps of course look beautiful. But I have no idea if they achieve your goal for SWAT modelling.

We reduced the number of figures and focused on less, but show the incredible detail as inset to demonstrate the high-resolution.

Figure 9 is clearly result evaluation. But you refer to it only in the methods 2.2.4. I like the plot but cannot read the axes. I think this plot is too small. Maybe you can think of another form to give this reference? Moreover, I am a little puzzled about the continuous histograms in SoilGrids but the very few distinct bars in your database for texture. Also the other parameters do not really appear to resemble SoilGrids. I am not at all saying that SoilGrids is more correct. But you really have to discuss

why you come up with different values and why your data is trustworthy.

We removed the pair-plot grids and focus on discussion and rather than 1:1 comparison as validation.

Discussion:
Maybe you could move some of the open issues from the methods section to the discussion? I would expect a little more structure what is discussed here.

We indeed removed parts from the introduction completely for focus and took a few sentences on soviet vs WRB/USDA particle size to discussion.

Again, I am sorry to return the manuscript before sending it to reviewers. I simply fear that we might enter an internal loop of rather structural deficits which obscure to dive to the actual matter of your work. I hope you can easily follow my suggestions and that you find it helpful. In case of any questions, please contact me. We can even quickly arrange a video-call.

All the best.
Conrad

On behalf of all authors,

Thank you very much for your guidance. We apologize for the hard-to-follow changes, as we have drastically restructured and re-written large sections of the manuscript, it is almost a new submission. We hope to have fulfilled your expectations and look forward to hearing back.

Alexander Kmoch

EstSoil-EH ~~– v1.0~~: A~~n~~ high-resolution eco-hydrological modelling parameters dataset ~~derived from the Soil Map of~~ for Estonia

Alexander Kmoch[1], Arno Kanal[1,†], Alar Astover[2], Ain Kull[1], Holger Virro[1], Aveliina Helm[3], Meelis Pärtel[3], Ivika Ostonen[1] and Evelyn Uuemaa[1]

[1]Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, Tartu, 51003, Estonia
[2]Chair of Soil Science, Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Fr.R. Kreutzwaldi 5, Tartu, 51014, Estonia
[3]Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, Tartu, 51005, Estonia
[†] deceased, 07th of May, 2019

Correspondence to: Alexander Kmoch (alexander.kmoch@ut.ee)

Abstract. To understand, model and predict landscape evolution, ~~functions~~ecosystem services and hydrological processes the availability of detailed observation-based soil data is extremely valuable. For the EstSoil-EH dataset, we synthesized more than 20 extended eco-hydrological variables for Estonia (Kmoch et al., 2019a; doi:10.5281/zenodo.3473289) as numerical and categorical values from the original Soil Map of Estonia, the Estonian 5m Lidar DEM, Estonian Topographic Database and EU-HydroSoilGrids layers.

The Soil Map of Estonia ~~is a vector dataset that~~ maps more than 750 000 soil units throughout Estonia at a scale of 1:10 000 and forms the basis for EstSoil-EH.~~.~~ It is the most detailed and information-rich dataset for soils in Estonia, with 75% of mapped units smaller than 4.0 ha, ~~and~~ based on Soviet era field ~~-~~mapping. ~~a Baltic country with an area of approximately 45 000 km².~~ For each soil unit, it describes the soil type (i.e. soil reference group), ~~quality,~~ soil texture, and layer information with a composite text code, ~~that~~ which comprises not only of the actual texture class, but also of ~~the~~ classifiers for ~~the~~ rock content, peat soils, ~~its~~ distinct compositional layers and their depths. ~~However,~~ ~~t~~To use these as~~it~~ eco-hydrological process properties in ~~as an input for numerical~~ modelling ~~using process-based physical models,~~ applications we translated the~~se~~ text codes ~~must be translated~~ into numbers. The derived parameters include soil profiles (e.g., layers, depths), texture (clay, silt, sand components), coarse fragments and rock content. In addition, we aggregated and predicted physical variables related to water and carbon (bulk density, hydraulic conductivity, organic carbon content, available water capacity). ~~Various generalisations and aggregations for agricultural soils for less-detailed versions of the map have been made at a scale of 1:100 000 and 1:200 000.~~

~~In this study, we create an extended eco-hydrological dataset for Estonia, the EstSoil-EH v1.0 (Kmoch et al., 2019a; doi:10.5281/zenodo.3473289), containing derived numerical values for the following data in all of the mapped soil units in the 1:10 000 soil map: soil profiles (e.g., layers, depths), texture (clay, silt, sand components), coarse fragments and rock content, and physical variables related to water and carbon (bulk density, hydraulic conductivity, organic carbon content). Ultimately,~~

~~our objective was to develop a reproducible method for deriving numerical values to support modelling and prediction of eco-hydrological processes in Estonia using the popular Soil and Water Assessment Tool.~~

The developed methodology and dataset will be an important resource for the Baltic region, but possibly also all other regions where detailed field-based soil mapping data is available. Countries like Lithuania and Latvia have similar historical soil records from the Soviet era that could be turned into value-added datasets such as the one we developed for Estonia.

1 Introduction

Soil has remarkable complexity and through its various functions plays ~~the~~a key role in ~~the~~ Earth's ecosystems and provides multiple ecosystem services to humans, such as food, and clean water. Recent studies have highlighted the role of properly functioning soils that can provide their ecosystem services for the achievement of the United Nations (UN) adopted the Sustainable Development Goals (SDGs) (Keesstra et al., 2018, 2016). Therefore an accurate quantitative description and prediction of soil processes and properties is essential in understanding the impacts of climate and land use changes on ecosystem services (Van Looy et al., 2017). For this purpose, spatially accurate maps of soil properties are needed, but unfortunately, these are either missing for many countries and regions in the world or ~~which~~ exist~~st~~ with insufficiently fine spatial resolution (Nussbaum et al., 2018). However, for many countries still ~~some~~ field-based data on soil properties are available. Moreover, the recent increase of spatial environmental data created by remote sensing (climatic, terrain variables etc.) can be ~~made~~ use~~fuld~~ for deriving the desired soil properties at fine resolution. There are several useful approaches to combine, or fuse, several datasets into one and obtain ~~full~~ desired complete spatial coverage of the soil properties needed for modelling.

At the global level, two main soil databases are available. The first was made available by the United Nations Food and Agricultural Organisation (FAO): the Harmonized World Soil Database (HWSD) v1.2 (Fischer et al., 2008). The dataset is a 30-arc-second ~~-~~raster database (approx. 100 ha resolution) with more than 15 000 different soil mapping units. It combines existing regional and national updates of soil information from around the world. Another global level soil ~~dataset~~database is SoilGrids250m (Hengl et al., 2017), which provides harmonized gridded soil data with values for sand, silt, clay, and rock fractions, and organic carbon and carbon stocks at several depths with a resolution of approx. 6.5 ha~~, which~~and can be used as inputs for eco-hydrological models e.g. SWAT. ~~SoilGrids also provides a harmonized soil database for Europe. The~~ SoilGrids250m has been derived with machine learning methods and using environmental variables, such as terrain properties, as predictor variables and field-based soil profiles as training set (Hengl et al., 2017). This approach takes advantage of recent abundance of high resolution environmental spatial data mostly obtained from remote sensing (e.g. terrain, climatic variables, soil moisture), and employs these datasets as explanatory variables to model soil properties at fine spatial resolution (Nussbaum et al., 2018). Analogously, EU-HydroSoilGrids (Tóth et al., 2017) provides a 3D soil hydraulic database for Europe based on SoilGrids250m and trained pedotransfer functions (PTFs). In other words, they~~it~~ use~~s~~ machine-learning regression as specialised ~~pedotransfer functions~~PTFs ~~(PTFs)~~.

PTFs are predictive functions of certain soil properties using data from field-based soil surveys (e.g. soil profiles). However, the potential of available PTFs has not fully been exploited and integrated into eco-hydrological modelling and ecosystem services provided by soils (Van Looy et al., 2017). For example, Soil organic carbon (SOC) is an important indicator of soil health and plays a key role in the global carbon cycle and therefore it is crucial to adequately quantify and monitor soil organic carbon (SOC) changes (Vitharana et al., 2017). However, reliable estimates for SOC have been difficult to obtain due to a lack of global data on the SOC content of each soil type (Eswaran et al., 1993). Very few SOC datasets are available for countries or regions. For example, the Northern Circumpolar Soil Carbon Database (Tarnocai et al., 2009) was developed to describe the SOC pools in soils of the northern circumpolar permafrost region. SOC stocks were also predicted under future climate and land cover change scenarios using a geostatistical model for predicting current and future SOC in Europe (Yigini and Panagos, 2016). Prévost (2004) described predictions of soil properties from the SOC content, and found that SOC was closely related to soil bulk density (BD) and porosity. Suuster et al. (2011) emphasized the importance of BD as an indicator of soil quality, site productivity, and soil compaction and proposed a PTF for the organic horizon in arable soils in Estonia. Abdelbaki (2018) evaluated the predictive accuracy of 48 published PTFs for predicting BD using State Soil Geographic (STATSGO) and Soil Survey Geographic (SSURGO) soil databases from the United States. They also proposed and validated a new PTF for predicting BD using SOC inputs.

However, these regional datasets are ~~not~~ often not detailed enough for ~~a~~ country~~-~~level applications nor do they benefit fully from ~~the~~ local high-resolution field-based soil data, as is the case for Estonia. There is no national scale dataset of measurements or predictions of SOC or bulk density (BD) for Estonia, and no large-scale high-resolution soil database is currently available with numerical data for a range of typical eco-hydrological process-based models. However, Estonia has a national ~~level~~ highly detailed digitized soil map (1~~-~~:~~-~~10 000) with 75% of mapped units smaller than four ha. It was ~~-~~created based on ~~Soviet-era~~ extensive ~~field-mapping~~field mapping during the Soviet-era ~~which~~and can serve~~s~~ as an excellent basis for PTFs and robust models ~~that can~~to predict soil properties at any given location (Minasny and Hartemink, 2011).

~~Eco-hydrological numerical models like the Soil and Water Assessment Tool (SWAT; https://swat.tamu.edu/) or the Regional Hydro-Ecologic Simulation System (RHESSys) have been developed and applied during the past 30 years to evaluate the effects of alternative management decisions on water resources and non-point-source pollution in river basins through the simulation of physical processes (Arnold et al., 1998; Douglas-Mankin et al., 2010). SWAT is widely used internationally and is increasingly applied in Northern European and Baltic watersheds to better assess the hydrological state of the environment based on modelling of the most relevant physical processes (Piniewski et al., 2018; Tamm et al., 2016, 2018). However, a main input factor for many of these models is detailed soil data, which does not exist for many countries on national scale or which exists with insufficiently fine spatial resolution. In addition, it is complicated to derive the values of the model parameters.~~

The objective of the present study was to develop a numerical soil database, EstSoil-EH, for modelling and for predicting eco-hydrological processes in Estonia and to ~~give~~provide a solid basis to estimate ecosystem services. The foundation of EstSoil-EH is the Soil Map of Estonia, which~~. The dataset~~ includes information about ~~for~~ soil profiles (e.g.,

layers, depths), textures (clay, silt, and sand contents), coarse fragments and rock content. We derived numerical values for the key characteristics for the whole of Estonia. ~~The h~~High-resolution environmental data available nowadays allow to develop improved PTFs, and modern advanced methods (e.g. machine learning, geostatistics) extrapolation and upscaling (Gunarathna et al., 2019; Van Looy et al., 2017). Thus, we employed machine learning and PTF-s to derive and aggregate additional physical variables related to the water and carbon cycle ~~these parameters~~ based on high-resolution field-based data and other environmental covariates (e.g. terrain variables).~~In this study, we derived numerical values for the following data in all of the mapped soil units in the 1:10 000 soil map: soil type (i.e. soil reference group), texture class, soil profiles (e.g., layers, depths), texture (clay, silt, sand components, and coarse fragments), rock content, and physical variables related to the water and carbon cycle (organic carbon content, bulk density, hydraulic conductivity, available water capacity and erodibility factor). We present also describe the development of a reproducible method for deriving numerical values from a the Soil Map of Estonia to support modelling and prediction of eco-hydrological processes with the popular Soil and Water Assessment Tool and we create provide an extended ready-to-use dataset containing the additional parameters.~~

2 Materials and Methods

2.1 Pre-processing ~~and screening of~~ the ~~initial~~base ~~soil~~ data~~base~~set

We performed extensive database standardisation on the original Soil Map of Estonia as the working basis and synthesise all further variables based on the standardised dataset sequentially. Figure 1 illustrates the major working packages and their in- and outputs of eco-hydrological parameters. The subsequent sections of the manuscript are structured accordingly: Section 2.2 Textural properties; Additional topographic variables, areal proportions of drainage and land use/land cover as predictor variables in section 2.3, section 2.4 describes the SOC RF model and BD PTF; and hydrological parameters added in section 2.5.

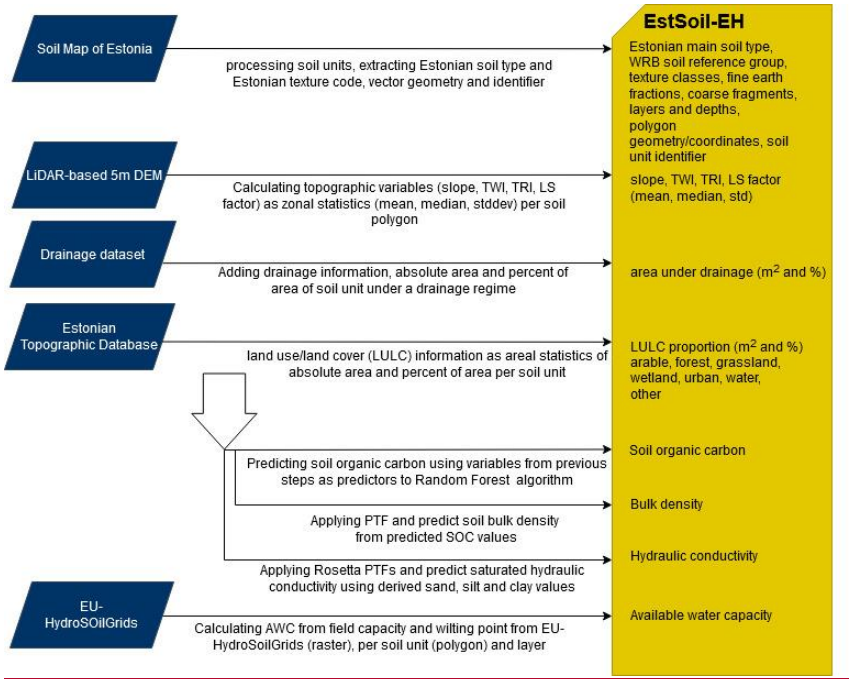

Figure 1: Flowchart for executed processing steps outlining the work packages, in- and outputs

The ~~source~~ base dataset ~~—~~ the original Soil Map of Estonia - ~~as described in the article, is not based on modelled but~~ is based on fully observed data (e.g. texture, soil profile depth, rockiness, presence of organic layer etc). Systematic mapping of Estonian soils to produce a paper-based soil map in scales 1:5 000 and 1:10 000 was started in 1954 (Reintam et al., 2005), with most intensive field studies in the period 1965-1969, for the main purposes of land evaluation and assessing potential for agricultural use. Generally, field mapping was carried out in scale 1:10 000 but in hilly or undulating areas with higher soil diversity in scale 1:5000, which resulted in mapped units with areas as small as 2500 m^2. ~~In~~ During 1982-1988 older mapping data was updated and new areas were included with full-area soil quality (primarily fertility, rockiness, water regime, texture, erodability) assessment. ~~In~~ During 1988-1990 soil field studies were performed for~~in~~ non-arable lands and ~~new mapping of~~ ameliorated lands. Forest soils were mapped ~~in~~ during ~~period~~ 1976-1989. During these large-scale ~~field mapping~~field-mapping ~~of soils~~activities, ~~the~~ soil texture was determined in situ based on organoleptic methods (feel methods) and for reference profiles laboratory analyses were performed. This enabled calibration between texture defined by

organoleptic method by each researcher participating in field survey and texture determined in the laboratory (Estonian Landboard, 2017). As athe result of large-scale soil mapping, 119 soil varieties in Estonian national classification system have been distinguished in the Estonian national classification system and more than 500 combinations of textural status have been described. About 10 ,000 profiles up to 1-meter depth (1 profile per 330ha) have been sampled and analysed for characterisation of mineral soils (Reintam et al., 2003, 2005). Thus, the texture codes and soil types assigned to the ca. 750 000 mapped soil units (polygons) are based on many decades of in-situ land surveying practices.

Between 1997 and 2001, the soil map was digitized and attribute data was inserted into the database, resulting in the official National Soil Map of Estonia, as a GIS vector dataset that mappedcontaining 750 000 soil units. It is at a scale of 1:10 000 (Estonian Landboard, 2017; https://geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Mullastiku-kaart-p33.html). The original Soil Map of Estonia is available vector layer for geographical information system software that can be downloaded from the Estonian Land Board (https://www.maaamet.ee/en) in several formats under a permissive open data license(Estonian Landboard, 2017). A copy with the original shapefile dataset, the related required documentation and checksums has been archived for reference (Estonian Landboard, 2017).

The Estonian soil map contains the following used attribute fields:

- Soil type: a designation of the soil name, the Estonian analogue to the WRB soil reference groups
- Texture: combination of texture classes defined for fine and coarse fragments, and to which depths the same texture and coarse fragments are observed (layer)

These attributes are encoded as “string” values, which include both letters and numbers. The important fields soil type and texture, are not just stored as standardised class values, but are instead a coded description based on abbreviations that are then combined with numbers for example depths and indictors for level of erosion, and are grouped together for different depths within the same attribute field. These description-based attribute values make it difficult to derive the foundational numerical values for sand, silt, clay and coarse fragments from the codes and to make them more consistent and usable in calculations and statistical analyses. In addition, our data screening revealed that the attribute values sometimes contradict the official legend for the Soil Map of Estonia. For example, the soil type reference sheet provided with the soil map lists ca. 120 different soil types in Estonia (Estonian Landboard, 2017; "muldade_tabel.pdf") and the soil legend document describes 9 main texture classes and 12 soil skeleton types, i.e., coarse fragments and rock morphology (Estonian Landboard, 2017; "mullalegend.pdf"). However, the database’s attribute table contains 7067 unique variations for soil types, which resulted from the use of many specific local derivatives and transcription errors. Similarly, the texture column actually contains 87240 unique values instead of 9, 21 (9+12) or 108 (9x12). Considering the possible permutations of these soil types and textures, it would be prohibitively difficult to develop any kind of reasonable standardisation for the soil parameters before cleaning and unifying the dataset. Therefore, we performed extensive database standardisation on the original Soil Map of Estonia as the working basis and derive all further variables based from the standardised dataset sequentially. illustrates the four major working packages to derive the desired eco-hydrological parameters. The subsequent four sections are structured accordingly: Section 2.2 represents

step A (Textural properties); step B (Additional topographic variables as predictor variables) in section 2.3, section 2.4 describes step C (SOC RF model and BD); and step D (Hydrological parameters) is described in section 2.5.

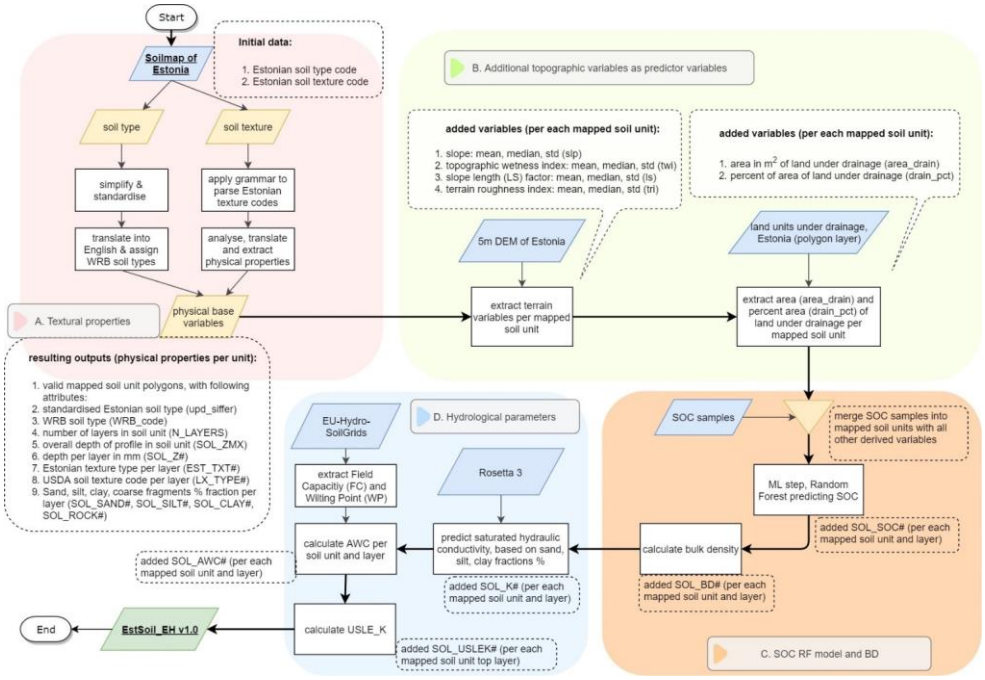

Figure 1: Flowchart for executed processing steps grouped into the four major work packages

2.2 Extraction of texture classes, soil reference groups, and deriving basic physical and textural values

The main implementation of this program is based on the Python library "Arpeggio" (Dejanović et al., 2016; https://pypi.org/project/Arpeggio/), which is a recursive-descent parser based on parsing expression grammars (also known as the Packrat parser; http://bford.info/packrat/). This let us express rules and symbols (i.e., the grammar) in such a way that our software could parse arbitrary text and find the various definitions of the texture in the same way as the rules are described in the map legend handout.

Listing 1 provides an example of a parsing grammar. At the start of the program, the basic elements are defined, starting with the 9 main fine-textured soil types: "*plsl, pl, tsl, tls, dk, sl, ls, s, l*". The parser honours the order of their

~~definition. Without these ordered rules, the parser will never find the more complex expression "*plsl*", because it would stop as soon as it encounters the "*pl*" part of the name. We also defined the coarse fragments types, and peat land soils.~~

~~The function "*def fine_textured(): return Optional(kPlus), fine-textured_list, Optional(amplifiers), Optional(depth_range)*" demonstrates the flexibility of how a parsing expression grammar parser can be configured. The function can find even optional (0 or 1) elements such as prefixes or suffixes within an arbitrary text stream.~~

~~Subsequently, several separators and special indicators must be defined that can precede or be appended in combinations to the abovementioned soil type elements. These were often formatted as subscripts or using special characters. This proved to be a major source of data-entry errors, encoding mistakes, and ambiguities, which had to be handled via additional error-checking code, e.g., lookup tables which are provided in the supplemental materials (Kmoch et al., 2019b; "soil_lib/LoimisLookups.py").~~

~~The mapped soil units also encode variations in the soil profile within a given soil unit. Thus, we must differentiate between a vertical separator for the observed soil layers, and a horizontal separator. However, we only considered the vertical component (soil horizons). In addition, these discrete vertical layers are only based the description in the original texture code. To capture and fully evaluate the possible texture codes, it was necessary to capture the meaning of any additional (rare) horizontal separators.~~

~~Because there are various data-entry errors and other ambiguities in the actual codes in the soil map dataset, we manually analysed all codes that could not be successfully evaluated by the grammar. Manual inspection was particularly required for codes that did not conform to the general rules described in the original soil legend handout. A full list of non-logical expressions, data-entry errors, and other grammar expressions that could not be easily or usefully standardised is provided as a supplemental Excel spreadsheet (Kmoch et al., 2019a; "texture_error_lookup.xlsx").~~

~~The parser for the defined grammar builds a data structure that can be evaluated for physical numerical parameters such as layers, depth, and the sand, silt, clay, and rock contents. This data structure is a Python dictionary object, i.e. a lookup table with nested key-value pairs that hold the parameters and the found values. In the example in Listing 2, it becomes apparent that there is a "/" vertical layer separator (at the bottom, the "code" parameter shows the original texture code for this soil unit), and that depths and fine fractions are accessible separately from the data structure. If a coarse fraction were defined in the texture code, then additional to the fine earth information, an additional "constituent" (the coarse fraction type) would be part of the respective layer (i.e. the "soilparts" object).~~

There are detailed studies on reference soil profiles in Estonia, Latvia and Lithuania that relate original soil texture, so called Katchinsky texture system (Kachinsky, 1965) to USDA soil system (Calhoun et al., 1998) and erosion modelling case studies where based on laboratory analyses transfer functions from Katchinsky to USDA texture classes were developed (Laas and Kull, 2003). The relationship between the Katchinsky and Atterberg systems were provided by R. Kask (2001).

The USDA soil taxonomy and World Reference Base soil classification systems use 12 textural classes, which are defined based on the sand, silt, and clay fractions (Ditzler et al., 2017). However, the USDA system defines fine particles as having a diameter ≤ 2 mm, whereas the Soviet-era maps use a diameter of ≤1 mm. The Soviet soil classification also mostly ignores the silt fractions, and focuses on the clay fraction (Ø ≤ 0.001 mm).

The Soil Map of Estonia's "texture" field encodes the texture and general soil layer structure for each mapped soil unit in a structured, rule-based format (based on old Soviet-era paper maps). The original observations were classified into the

Estonian texture code system based on Katchinsky (1965) soil particle size standards at the time of observation (not by us). Regarding uncertainties related to that process - as we take these as observed data - achieving 5% accuracy in organoleptic determination of clay content for lower value classes while possible error increased in case of heavy texture classes.

We developed a computer program that converts the encoded texture codes into an intermediate data structure, which extracts again Estonian texture classes, coarse fragment classes, into separate layers including depth information. Subsequently, we derived all defined numerical texture values using a lookup table (Table 1) that represents our best efforts to account for the size difference between the USDA and Soviet systems and lack of silt data in the Soviet system. The foundational numerical values for fine earth and coarse fragments fractions of soil are solely derived from the extracted processed ~~and translated~~Estonian texture classes as demonstrated in Table 1.

In additional we introduced two more classes beyond the well-known USDA textures classes, i.e. “PEAT” and “GRAVELS”. The former states that this soil unit is a peatland, where the peat layer thickness is at least 30 cm. For hydrological modelling reasons we decided to still assign sand, silt and clay fractions to these units in order to provide a continuous hydrological soil surface. To soil units with the class “PEAT” a high clay content was assigned in order to represent the low vertical conductivity at the bottom theses peat bogs. However, for applications that critically evaluate clay content for soil units, the additional “PEAT” texture class can be used to apply additional rules to mask these soil units accordingly. The latter class “GRAVELS” is intended to demark soil units or discrete layers therein, where only a coarse fragment type but no fine textures have been coded in the original texture codes. In these cases, depending on the type of the coarse fragment the layer can consist of gravels, large rocks or massive rock.

Table 1: Example of the basic rules for deriving numerical values for texture (sand, silt, and clay contents) from the Estonian texture codes and assigned new English and USDA texture classes. These rules were selected by the authors. The full table is provided as a supplemental Excel spreadsheet (“texture_rules_lookup.xlsx”)

Estonian texture code	Estonian name	English name	USDA texture code	Proportion (%) of total weight		
				Sand	Silt	Clay
l	Liiv	sand	S	90	5	5
l_1	sõre liiv	coarse sand	S	95	5	0
l_2	sidus liiv	fine sand	S	90	3	7
sl	saviliiv	loamy sand	LS	82	9	9
sl_1	saviliiv	loamy sand	LS	82	9	9
ls	liivsavi	loam	L	55	30	15
ls_1	kerge liivsavi	sandy loam	SL	65	20	15
ls_2	keskmine liivsavi	loam	L	55	30	15
s	Savi	clay	C	25	30	45

Similar to the Estonian texture classes there exist Estonian stoniness classes that describe a certain type of coarse fragments within the soil profile. An additional number in connection with this rock type identifier indicates the amount/volume of these rocks in 1kg of soil. We used this indicator number to designate numerical values for the coarse fragments. Table 2 shows how we derived the rock content from the coarse fragments indicator that we obtained from the soil map encoding.

A base assumption is that most soils in Estonia were sampled to a depth of 1 m, as this is the case for a default soil profile. There is only one vertical profile defined per mapped soil unit. If larger or smaller depth information was encoded in the original soil texture code, then this would be used for the overall depth of that soil sample. For each of the layers, we collated depth from the soil surface to the bottom of each layer. The data model is relational and each soil unit is represented as one row, with its polygon geometry, identifier and all collected parameters as attributes. The maximum number of distinctly defined layers was 4, and layer dependent parameter values (at different depths) are only meaningful where the variable's number suffix is smaller or equal the number of defined layers.

Table 2: The relationship between the coarse fragments (rock content and shape) indicator from the soil map encoding and the rock content as a % of the total volume. We used the average of each defined range.

	Scale of conversion for rock content					
"Skeleton" indicator number	1	2	3	4	5	6
Inferred rock content (% of volume)	6	15	25	40	60	85

We ~~in Python to find the best match from the soil type reference list (Kmoch et al., 2019b; "01_soilmap_soiltypes_textures_layers.ipynb"). The algorithm progressively shortens the name from the right and~~compared the encoded ~~results~~ Estonian~~with the~~ soil type from the original soil map in order to find the most appropriate soil type from the main Estonian soil types from the soil reference list. The soil types and the Estonian soil names were then related to the FAO World Reference Base (WRB) soil reference groups (FAO, 2015) after the data have been corrected and standardised for each map unit in the extended soil dataset based on expert input (Hiederer et al., 2011). The full table that relates the Estonian and WRB soil reference groups is provided with the supplemental materials.

This first and fundamental step concluded with a set of variables for each mapped soil unit that include now separate standardised Estonian and English/USDA texture classes per soil layer, number and depths of layers of the mapped soil unit and numerical values for fine earth and coarse fragments fractions per layer, and a WRB soil designation group.

2.3 Adding topographic variables as predictor variables

For the subsequent step of SOC prediction via the Random Forest machine-learning model, we calculated mean, median and standard deviation of several topographic and environmental variables as additional predictor variables. Topographic variables slope, Topographic Wetness Index (TWI), Terrain Ruggedness Index (TRI), and LS-factor were all calculated by using SAGA-GIS software based on a digital elevation model (Conrad et al., 2015). The LiDAR-based Digital Elevation Model with resolution ~~1~~5 m was obtained from Estonian Land Board.

~~2.3.1 Topographic Wetness Index (TWI)~~

The TWI is a topo-hydrological factor proposed by Beven and Kirkby (Beven and Kirkby, 1979) and is often used to quantify topographic control on hydrological processes (Michielsen et al., 2016; Uuemaa et al., 2018) which also are relevant in the soil evolution. TWI controls the spatial pattern of saturated areas which directly affect hydrological processes at the watershed scale. Manual mapping of soil moisture patterns is often labour-intensive, costly, and not feasible at large scales. TWI provides an alternative for understanding the spatial pattern of wetness of the soil (Mokarram et al., 2015). It is a function of both the slope and the upstream contributing area:

$$TWI = \ln\left({}^{a}/_{\tan b}\right) \qquad (1)$$

where a is the specific upslope area draining through a certain point per unit contour length ($m^2\ m^{-1}$), and b is the slope gradient (in degrees).

~~2.3.2 Terrain Ruggedness Index (TRI)~~

TRI reflects the soil erosion processes and surface storage capacity which again is relevant from a soil evolution perspective. The TRI expresses the amount of elevation difference between neighbouring cells, where the differences between the focal cell and eight neighbouring cells are calculated:

$$TRI = Y[\sum(x_{ij} - x_{00})^2]^{1/2} \qquad (2)$$

where x_{ij} is the elevation of each neighbour cell to cell (0,0). Flat areas have a value of zero, while mountain areas with steep ridges have positive values.

~~2.3.3 LS-factor~~

The potential erosion in catchments can be evaluated using LS factor as used by the Universal Soil Loss Equation (USLE). The LS factor is length-slope factor that accounts for the effects of topography on erosion and is based on slope and specific catchment area (as substitute for slope length). In SAGA-GIS the calculation is based on (Moore et al., 1991):

$$LS = (n+1)\left({}^{A_s}/_{22.13}\right)^n\left({}^{\sin\beta}/_{0.0896}\right)^m \qquad (3)$$

where n=0.4 and m=1.3.

~~2.3.4 Drainage area per mapped soil unit~~

In addition, we calculated the area per mapped soil unit in m^2 and in percent of area, which is under drainage. The drainage regimen considered both underground tile drainage and ditch based drainage systems. Analogously, we summarised land use/land cover proportions into arable land, forest, grasslands, wetland, urban areas, water and other also as area per mapped soil unit in m^2 and in percent of area. The drainage and land use/land cover information ~~used for this was compiled based on the~~was derived from the ~~S~~Estonian Topographic Data ~~Set~~base (ETAK) and the official register of drainage systems by the Agricultural Board of Ministry of Rural Affairs of Estonia.~~All the variables were calculated using the GIS software packages QGIS and SAGA.~~

2.4 Predicting Soil Organic Carbon (SOC) and Bulk Density (BD)

The main information retrievable from the Soil Map of Estonia are only the soil type and the soil texture. However, soil hydraulic properties and SOC data are needed for many different applications in soil hydrology, ecology and eco system services modelling. Pedotransfer functions (PTFs) have proven to be useful to indirectly estimate these parameters from more easily obtainable soil data (Van Looy et al., 2017). Therefore, several soil parameters like soil organic carbon, bulk density and saturated hydraulic conductivity must be derived via PTFs and other data assimilation methods. To apply PTFs and other data-assimilation methods, third-party datasets can be used as secondary sources. In the previous steps we have prepared a wide set of input variables, including the numerical fractions for the textural properties, standardised classes for soil type and soil textures, and additional topographic variables, which we can apply as predictor variables to model the value distribution for SOC and BD. We develop these two extended soil physical input parameters as organic carbon content in % soil weight (SOL_CBN# layer 1-4), and dry bulk density in Mg/m^3 or g/cm^3 (SOL_BD# layer 1-4).

In order to map the spatial distribution of SOC in Estonia a machine learning model random forest (RF) was used to predict SOC based on parameters derived from the soil map. RF was preferred to more advanced ML algorithms (e.g., neural networks) because it has shown to be relatively resilient towards data noise and not require preliminary hyperparameter tuning (Breiman, 2001; Caruana and Niculescu-Mizil, 2006). In addition, feature importances can be extracted from the model to determine the most influential predictor variables.

For training, we used measurements of soil organic matter (SOM) or soil organic carbon (SOC) from forest areas (samples sizes: n=100), 4 datasets of samples from Estonian open and overgrown alvars and grasslands (n: 94, 137, 146, 69), peatlands (n=175) and from arable soils transects (n=8964) resulting in 3373 distinct point locations (Kriiska et al., 2019; Noreika et al., 2019; Suuster et al., 2011). Where necessary, the SOM values were translated into SOC via: $SOC = SOM / 1.724$. Many samples from peatlands and arable fields were often sampled within the same mapped soil unit. For these soil units (polygons) the respective soil measurement data was averaged and joined to the respective soil units to reduce the bias of the prediction. After joining the sample size reduced to the 397 distinct training samples for machine learning (Figure 2).

This data was then randomly split into training (60%) and test (40%) sets and the model was evaluated by predicting SOC based on the predictor variables of the test set. Finally, the model was applied to soil map polygons without available SOC measurements to predict SOC content in Estonian soils.

Figure 2: Distinct soil unit polygons including all sampling locations for the ML training sample.

Subsequently, we calculated soil bulk density based on ~~texture values and~~ predicted soil organic carbon for each layer in each mapped soil unit polygon, with following PTF (Adams, 1973; Kauer et al., 2019), which has been successfully applied in Estonia:

$$BD = 1 / (0.03476 \times SOM + 0.6098) \quad (4)$$

where: $SOM = SOC \times 1.724$

The conversion factor of 1.72 is a widely used universal value. However, we acknowledge that the real value varies slightly between soils.

2.5 Assimilation of additional hydrological variables

In order for this dataset to be more useful in eco-hydrological modelling we developed and added two additional hydrological variables. Saturated hydraulic conductivity (K_{sat}) relates soil texture to a hydraulic gradient and is quantitative measure of water movement through a saturated soil. In addition to the ability of transmitting water along a hydraulic gradient we also add available water capacity (AWC) as a variable. AWC describes the soil's ability to hold water and quantifies how much of that water is available for plants to grow. We develop two variables saturated hydraulic conductivity (mm/hr), and available water capacity of the soil layer (mm H_2O/mm soil). We calculated K_{sat} using the improved Rosetta3 software, which implements a pedotransfer model with improved estimates of hydraulic parameter distributions (Zhang and Schaap, 2017). It is based on an artificial neural network (ANN) for the estimation of water retention parameters, saturated hydraulic conductivity, and their uncertainties. For each standardised texture class, we used the numerical fine earth fractions for sand, silt and clay as inputs for the Rosetta3 software and calculated K_{sat} for each layer in each mapped soil unit polygon. Table 3 demonstrates the predicted values for several texture classes.

Table 3: Predicted K sat values and reported standard deviation from Rosetta3

texture class	sand	silt	clay	k_sat	k_sat_std
GRAVELS	100	0	0	645.68	1.29
S (coarse sand)	95	5	0	362.25	1.19
S (sand)	90	5	5	133.21	1.13
S (fine sand)	90	3	7	113.71	1.17
LS (Estonian 'sl1-3' classes)	82	9	9	37.54	1.15
LS (Estonian 'tsl1' class)	80	14	6	40.2	1.18
SL	65	20	15	11.02	1.18
L (Estonian 'ls2' class)	55	30	15	9.04	1.21
CL	50	15	35	3.67	1.3
L (Estonian 'tls1' class)	40	45	15	8.16	1.37
SiL	35	50	15	8.89	1.35
SiCL	30	40	30	3.97	1.34
PEAT (Estonian 't1' class)	25	25	50	5.09	1.53
HUMUS	25	25	50	5.09	1.53
HC	25	30	45	4.29	1.43
C	25	30	45	4.29	1.43
PEAT (Estonian 't2' class)	20	20	60	7.24	1.81
PEAT (Estonian 't3' class)	15	15	70	9.2	2.45

~~(Kmoch et al., 2019b; "Rosetta 3.0").~~

In order to calculate available water capacity**,** we summarized the field capacity (FC, at −330 cm matric potential −0.03 MPa) and wilting point (WP, at −15,848 cm matric potential −1.5 MPa) variables of the 7 soil depths of the EU-

SoilHydroGrids 250m resolution raster datasets (Tóth et al., 2017) for each mapped soil unit for the provided depths of 0, 5, 15, 30, 60, 100, and 200 cm. The available water capacity is then calculated for each of the 7 depths by a raster calculation: AWC = FC - WP (Dipak and Abhijit, 2005). The resulting 7 AWC raster layers are then averaged into the respective depth ranges for each of the discrete layers of the Estonian mapped soil units. ~~The Python code of the process for the extraction of FC and WP from the EU-SoilHydroGrids is provided with the supplemental materials (Kmoch et al., 2019b; "05_hydrogrids_extents_and_awc_extract.ipynb").~~

3 Results

In this study, we developed the EstSoil-EH database, which includes standardised soil type and soil texture data from the official Soil Map of Estonia, related to the World Reference Base and FAO soil classes and USDA texture descriptions. Figure 3 shows a map of the classified topsoil texture classes derived from the original Estonian texture codes. In addition, it shows the peat soils that cover up to 20% of Estonia, and are an important soil type in such northern countries.

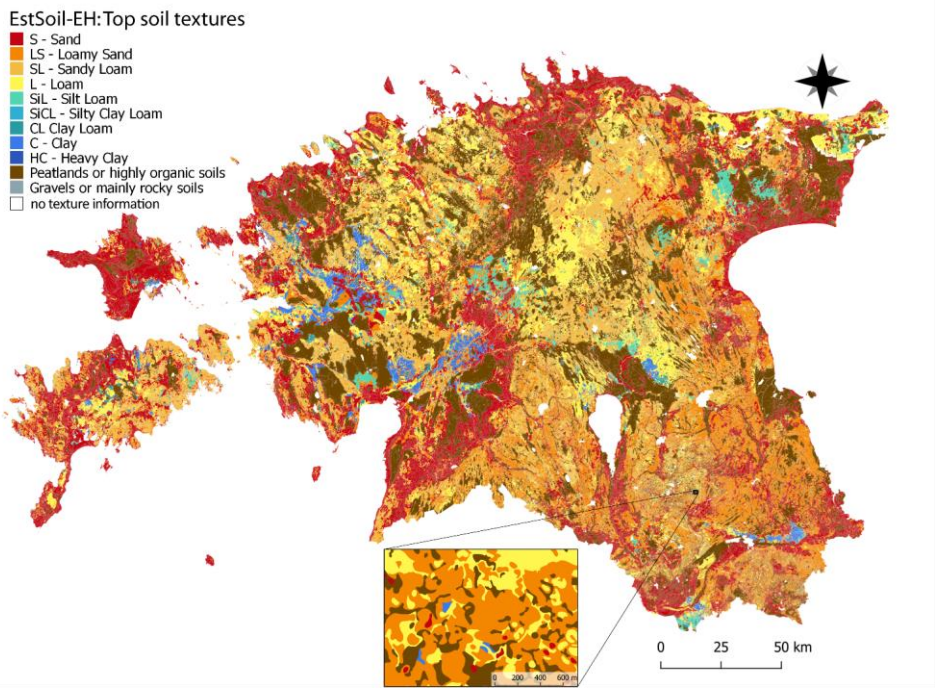

Figure 3: EstSoil-EH dataset: USDA topsoil textures derived from the original Estonian texture codes by the software developed in the present study, including additional classes "PEAT" and "GRAVELS". Lower image is a zoom-in to a small region to visualize the high-level of detail.

We synthesised additional information usable in an eco-hydrological modelling context for each of the soil units. These values include the number of discretized soil layers – up to a maximum of 4 separate vertical distinct soil layers where described in the original texture codes –the depth of each layer, and the maximum depth of the sampled profile for each mapped soil unit. Based on the layer information and the extract texture classes we defined the percent fractions per volume of sand, silt, clay, and coarse fragments per layer. We also added topographical and land use/land cover information and used these as predictors for soil organic carbon. Subsequently, we predicted bulk density and hydraulic conductivity and for each soil polygon's defined layers and assimilated available water capacity using inputs from EU-HydroSoilGrids. Table 4 contains the full list of variable and parameters per mapped soil unit contained in the EstSoil-EH dataset.

Table 4: Description of variables and parameters available in the EstSoil-EH dataset

name of variable per mapped soil unit	data_type	description
est_soiltype	string	Estonian soil type
wrb_code	string	FAO WRB soil ~~type~~ reference group (1st and 2nd level)
wrb_main	string	FAO WRB main soil ~~type~~ reference group (1st level)
est_txcode	string	reconstructed error-free interpretation of Estonian texture encoding description
nlayers	number	number of recognized layers/horizons
zmx	float64	depth in mm: max depth of the sample analysed soil profile in the mapped soil unit
z1-4	float64	depth in mm: the bottom of layer # (if nlayers indicates defined)
est_txt1-4	string	Estonian texture class per layer #
lxtype1-4	string	USDA texture class
est_crs1-4	string	Estonian coarse fragment type
sand1-4	int64	% mass of Sand in fine earth fraction
silt1-4	int64	% mass of Silt in fine earth fraction
clay1-4	int64	% mass of Clay in fine earth fraction
rock1-4	int64	% volumetric in kg soil
soc1-4	float64	% soil weight
bd1-4	float64	g/cm^3
k1-4	float64	mm/hr
awc1-4	float64	mm H_2O/mm soil
slp_mean	float64	mean slope, from DEM (also median and stddev)
twi_mean	float64	mean terrain wetness index (also median and stddev)
ls_mean	float64	ls-factor, (also median and stddev)
tri_mean	float64	terrain roughness index, (also median and stddev)
area_drain	float64	area per unit under a (e.g. tile-)drainage regimen
drain_pct	float64	percent of the area of the soil unit under drainage
area_arable	float64	area m^2 of LULC arable (6 add. LULC types)
arable_pct	float64	% of area is LULC arable (6 add. LULC types)
geometry	geometry	polygon, EPSG:3301 Estonian National Grid

3.1 Validation of soil type and texture classes extraction and standardisation

For the main soil types, we achieved 97.7% agreement between the software's result and the manual classification. The manual verification of the validation revealed several re-labelling issues from the error lookup table. A visual assessment by two soil sciences senior research staff asserted that the level of similarity of the soil types that were selected by the automated

process were closely related. However, the mismatches (1943 records, equivalent to 2.3% of the total records) indicated that the soil experts tended to interpret "errors" based on personal knowledge that may not be reproducible in a strictly automated fashion. For example, some landforms (e.g. eroded material filling low slopes or collapsed cliffs) were originally classified as exceptions to the general classification rule based on the local knowledge of the landscape. When standardising these expert interpretations with the same more general soil type, we reduced the number of mismatched soil type identifiers to 0. Furthermore, it should be emphasised that humans tend to make mistakes when performing repetitive procedures. Therefore, we consider the high accuracy (97.7%) to be a very good result.

For the validation of textures, we used several steps. First, given the high agreement between the software-generated codes and the human-generated codes, we accepted the software's texture codes for use in our subsequent evaluations. Next, we compared the extracted main texture for each layer with the manually coded value:

- 77 870 of 83 364 records (93.4%) showed identical parsing of the full texture code
- 71 635 of the records (85.9%) showed identical interpretation of the first layer's texture type (10 312 records were differently coded, and 1417 produced "no value" errors, in which either the source or validation dataset contained no value, preventing a comparison with the other dataset's value)
- 65 000 of the records (78.0%) showed identical interpretation of the second layer's texture (with 2325 differently coded textures, and 16 038 "no value" errors, of which 15 461 occurred in the automated processed new dataset, and only 577 occurred in the validation dataset)
- 82 507 of the records (99.0%) showed identical interpretation of the third layer's texture (with most errors caused by a non-existent third layer, 334 differently coded, and 523 with a "no value" error)

For sand, silt and clay fractions we could obtain laboratory analysis **only for forest soil samples**. We calculated the root mean squared error (RMSE) and chose the Normalized Median Absolute Deviation (nMAD) as an additional measure of dispersion of error for non-gaussian distributed data:

- RMSE for sand: 13.1 %
- nMAD for sand: 9.68
- RMSE for silt: 10.7 %
- nMAD for silt: 7.0
- RMSE for clay: 6.5 %
- nMAD for clay: 3.9

Our manual assessment of the mismatches indicated the same problem that occurred with the soil types. The expert assessments aimed to keep as much information as possible available in their decoded classification, and this did not always agree with the automated processing rules. It is not possible to retrospectively redefine minor differences in boundaries between different classes between texture systems, but we consider natural variation of texture within the soil mapping unit in scale 1:10 000 more significant than that of different texture systems. Furthermore, the complexity of the Estonian texture rules and

the reliance on human judgement creates high uncertainty in some cases, even for human interpretation. In addition, to derive the grammar rules, we added a few simplifying elements, such as omitting some rarely used additional information in the soil texture descriptions. For example, the Estonian rules allow specification of several soil parts, but as a horizontal distribution within the same mapped soil unit rather than as vertical layers. This is understandably complex, making it difficult to classify this variable soil as a single soil unit. Consequently, it is inevitable that some of these descriptions will not agree with the software's classification.

3.2 SOC prediction and validation of Random Forest model

We also calculated several extended soil properties, i.e. *SOC* content and *BD*. The RF regression model was implemented with the RandomForestRegressor function from the Scikit-learn Python library. The model was evaluated by predicting SOC based on the predictor variables of the test set for the 60:40 split. Figure 4 illustrates the cross-validation scatterplots of observed vs. predicted SOC values for the test/validation sample splits. Following characteristics are reported for the chosen RF model:

- coefficient of determination (R^2) score: 0.69
- score of the training dataset with out-of-bag estimate (oob score): 0.58
- Pearsons *r correlation coefficient*, training: 0.90, validation: 0.83

RF feature importances, top 6:

- Clay content (SOL_CLAY1): 0.65
- Terrain Roughness Index, standard deviation (tri_stdev): 0.04
- Sand content (SOL_SAND1): 0.03
- LS-factor, median (ls_median): 0.028
- Area under drainage in percent (drain_prct): 0.027
- Coarse fragments rock content (SOL_ROCK1): 0.024

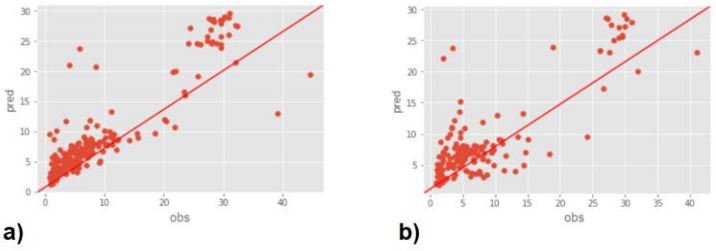

Figure 4: Random Forest model cross-validation scatterplot of observed vs. predicted SOC values for the test/validation sample splits: a) training subsample and b) validation subsample

Figure 5 shows the predicted values of SOC for the top layer. On visual inspection the spatial distribution for the SOC content matches comparatively well with known agricultural areas, where low carbon content prevails, as well as with the peat land areas, which have a very high carbon content.

For further description and guidance on errors in the predictions for SOC and BD we calculated the RMSE and nMAD as an additional measure of dispersion of error for non-gaussian distributed data. BD observed data was only available for arable lands and forest soil samples, and should be treated accordingly.

- RMSE for SOC predictions: 2.95 %
- nMAD for SOC: 1.44
- RMSE for the subsequent BD predictions with PTF: 0.33 g/cm³
- Normalized Median Absolute Deviation (nMAD) for BD: 0.15

However, due to the small number and distribution of input samples over four distinct land cover types, namely arable lands, wetlands, forests and open/grass lands, we broke down the error distribution for these for land forms in Table 5. The prediction error characteristics differ, with the smallest errors for arable lands, then wetlands and the largest errors for open grasslands and forest.

Table 5: ~~Table 6:~~ Statistical description of SOC prediction error per land form

landform	mean	std	min	25%	50%	75%	95%	max	median	nMAD	kurtosis	skew	RMSE
wetland	1.74	2.73	-5.22	-0.05	1.71	3.51	6.66	8.09	1.71	2.15	-0.1	0.04	3.23
arable	-1.54	1.78	-21.2	-2.12	-1	-0.63	-0.12	6.82	-1	1.12	29.65	-4	2.35
forest	-2.08	4.46	-24.56	-3.07	-1.52	-0.09	3.44	25.65	-1.52	2.79	7.39	-0.44	4.92
grassland	1.06	4.28	-8.47	-1.79	0.52	2.92	10.94	11.78	0.52	3.21	1.16	0.59	4.38

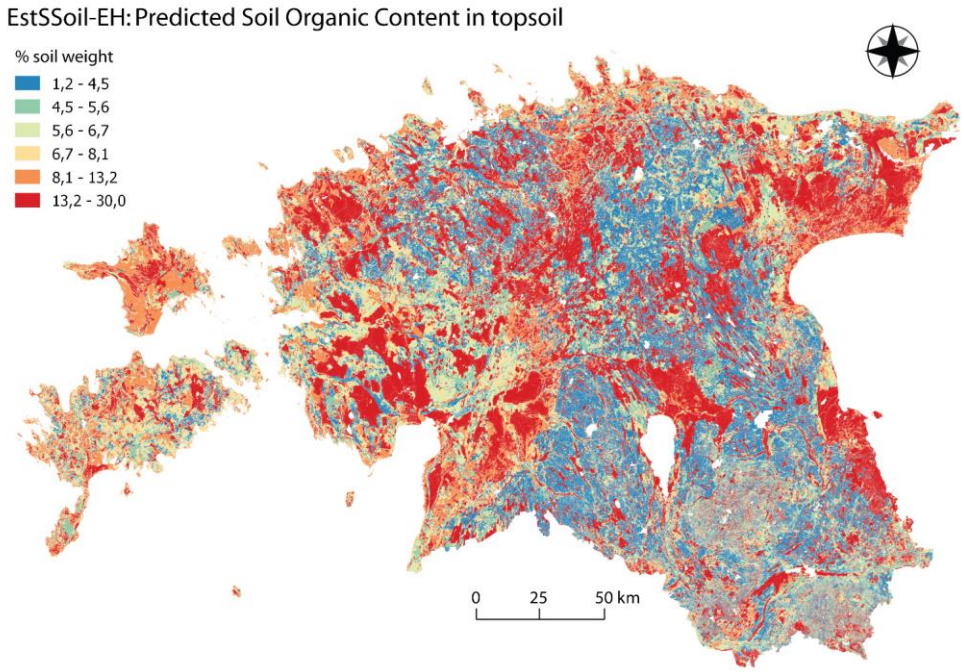

Figure 5: ~~**Extended physical soil parameters: a) predicted**~~**Predicted soil organic carbon (SOC)** ~~**and b) bulk density (BD) of the first layer.**~~**of the top soil layer**

3.3 ~~Extended~~ Hydrological variables results

Based on the variables derived in previous steps, we could ~~also~~ calculate ~~soil hydraulic parameters, such~~ saturated hydraulic conductivity (K_{sat}) based on the sand, silt and clay content. ~~and a~~Rosetta reports the standard deviation for its internal prediction process, which draws many samples for the same input of sand, silt and clay content and then provides the mean as the predicted value for K. The summary of ~~T~~the predicted K_{sat} values and the standard deviation were summarized in Table 3. ~~f~~For peat areas and wetlands the predicted values also corresponds with ranges reported in the literature for the sand-silt-clay ratios provided (Gafni et al., 2011).

Available water capacity (AWC) was calculated solely by aggregating EU-SoilHydroGrids data for field capacity and wilting point (Tóth et al., 2017). ~~The USLE *K* erodibility factor for each soil unit was also calculated (Figure 8).~~ We compiled

all parameters into a dataset that can now be easily used with SWAT or other eco-hydrological and land-use-change models. As we are not changing the general geometry or underlying spatial data model of the original soil map, all parameters are only added to the existing mapped soil units and thus, all original soil polygons remain discernible. ~~The dataset (Kmoch et al., 2019a; doi:10.5281/ZENODO.3473290) and the software codes (Kmoch et al., 2019b; https://zenodo.org/record/3473210) have been deposited online.~~

4 Discussion and Future Work

For EstSoil-EH, we derived numerical values for the following data in all of the mapped soil units in soil map: soil type (i.e. soil reference group), texture class, soil profiles (e.g., layers, depths), texture (clay, silt, sand components, and coarse fragments), rock content, and physical variables related to the water and carbon cycle (organic carbon content, bulk density, hydraulic conductivity, available water capacity). Before our analysis, a large amount of the information of the high-resolution Soil Map of Estonian was not readily usable beyond the field or farm-scale because of the need to manually interpret the specialised soil types and the complexity of the rules that describe the texture or other characteristics of the soil units. We also describe the development of a reproducible method for deriving numerical values from a national survey-based soil map to support modelling and prediction of eco-hydrological processes, and ecosystem services, and we provide an extended ready-to-use dataset containing additional parameters. ~~It is widely used in Estonia. But before~~The presented~~developed dataset is of very high spatial detail based on the original Estonian national soil map, which was created from directly surveying all of Estonia.~~ Thus, our presented dataset holds to potential to further improve our understanding of eco-hydrological processes in the landscape through the use of advanced ~~numerical~~ statistical (e.g. machine learning) and process-based models. The derived information is much more spatially related to the landform/landuse observed there than any other dataset covering soil information for Estonia. Furthermore, the textures and SOC/BD values are directly derived from reliable observed data samples from Estonia, with a reproducible workflow, which is unique in the case of Estonian soil datasets, whereas this is not true for many other reported soil datasets that cover the area of Estonia. Furthermore, the open access availability and transparency of measurement data can provide a reliable building block for advancing studying soil and hydrological processes in Estonia into temporal aspects. ~~t~~Especially, properties such as SOC and BD will vary extensively depending on the land use and land cover. In combination which developments that capture also the dynamics of land use change and adaptations under climate change, the evolution of the soils in Estonia could more readily be investigated.

One challenge in SOC modelling was that the number of field-based samples that was used for training the Random Forest model was relatively small for the whole country. Even though the samples covered four main land cover types (agricultural, forests, wetlands, grasslands), there was still significant spatial heterogeneity that might have not been captured. Moreover, in addition to field-based validation data, we used lower resolution modelled datasets, e.g. SoilGrids and EU-SoilHydroGrids, for a comparative validation. These datasets are not necessarily more accurate than the results of our classification. Although we accounted for this problem by providing additional comparisons, the scale mismatch between

continuous raster datasets and polygon-based data inevitably introduced errors and trade-offs into the comparison. One solution to these problems would be to perform supplemental field sampling to ground-truth the source data and confirm the accuracy of our model's classification based on the field data.

~~A direct interpretation on the derived discrete layer information as soil horizons needs should not be generalized but checked on per case basis.~~ From the point of end-user, the first layer is not a default 30 cm deep top soil layer. A direct interpretation of the derived discrete layer information as soil horizons should not be generalized but checked on per case basis. All physical, chemical and hydraulic properties are based on the analysis of the original texture code per mapped soil units and the resulting discrete layers per unit. This is an important usage constraint, for example in sense of biological activity, as 30 cm soil layer is most active, but for each soil unit it needs to be checked which layers extend into which actual depths. Also the *SOC* content and *BD* are not modelled in a vertical continuum but per discrete values per unit and layer. However, fertile soils, like Luvisols contain a lot of *SOC* also in deeper layers. But such additional expert knowledge is not encoded in original Soilmap of Estonia, nor in the processing algorithms that derived the extended parameters for this newly generated dataset. However, such additional knowledge, as well as more appropriate models for peatland areas, could be included as additional rules in a subsequent improvement of this dataset.

Kõlli et al. (Kõlli et al., 2009) published estimates of the SOC stocks for forests, arable lands, and grasslands and for all of Estonia. Nevertheless, they constrained their finding by noting that their estimates were calculated based on the mean SOC stock for each soil type and the corresponding area in which the soil type was distributed. Putku (2016) used the large-scale Soil Map of Estonia at the polygon level for SOC stock modelling for mineral soils in arable land of Tartu county. Carbon content calculations in Estonia have historically been predominantly made for soils in agricultural areas. Exisiting literature and our results in summary are in line with SOC distribution per soil type in mineral soils in arable lands (E. Suuster et al. 2012).

The original purpose of this dataset was to derive values for hydrological modelling purposes and at the same time to stay as close to the original data as possible. From that perspective peat soil units are currently modelled with assumptions to have a similar behaviour to clay hydrologically. Therefore, the spatial distribution of clay percentage in particular, but also the concurrent physical fractions of sand and silt do not make scientific sense for these areas where peat is prevalent. In order to make the dataset as useful as possible and to identify peatland areas, we introduced the additional class "PEAT" into the USDA classification. While sand, silt, clay and rock content are directly derived values from the original texture codes, *SOC* and K_{sat} are modelled via statistical machine-learning algorithms, which include additional uncertainty. This should be considered when evaluating *BD* ~~and USLE *K*~~, which are calculated using *SOC* as an input variable. In addition, it would be possible to use *BD* as an additional predictor for Rosetta. However, we decided that this would introduce too much uncertainty as *BD* in EstSoil-EH is based on a PTF function of *SOC*, which in return was also predicted via statistical modelling.

The only variable which we did not model based in dependence of already modelled parameters was *AWC*. Here we summarised the EU-SoilHydroGrids 250m (Tóth et al., 2017) raster datasets for *FC* and *WP* as inputs an external data

integration. This is not ideal and can be considered a trade-off between introducing too much uncertainty and an external un-related data source.

In the future, we foresee step-wise improvement of our software by developing better PTFs to estimate parameters and to better integrate the presence of peat soils and other specific landscapes and environments in Estonia. Furthermore, statistical machine-learning or neural network and deep learning methods could be tested in order to improve soil classifications and express more complex relationships between soil types and textures. Currently, one specificity of the newly created EstSoil-EH dataset is its discrete nature, as we are only adding derived numerical variables to the existing mapped soil units (polygons). We do not predict a continuous surface in this study, thus, comparisons with continuous surface parameters predicitons such as in SoilGrids (Hengl et al., 2017) or EU-SoilHydroGrids (Tóth et al., 2017), are not directly possible. However, the workflow could potentially be extended also for creating continuous surface. With appropriate modification (e.g., to use the soil characteristic codes more consistently for a different country), our methodology could also be applied in other countries such as Lithuania or Latvia that share similar historical land- and soil surveying practices.

Code and data availability.

The described "EstSoil-EH" dataset including all supplemental tables and figures is deposited on Zenodo, doi:10.5281/ZENODO. 3473289 (Kmoch et al., 2019a). Supplemental software and codes that were used, e.g. the texture-code parsing scripts, the machine learning model and the parameter calculation Jupyter notebooks are maintained on GitHub (https://github.com/LandscapeGeoinformatics/EstSoil-EH_sw_supplement/releases) and were also deposited on Zenodo, doi: 10.5281/zenodo.3473209 (Kmoch et al., 2019b). The original National Soil Map of Estonia (https://geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Mullastiku-kaart-p33.html) was archived for reference on the DataCite- and OpenAire-enabled repository of the University of Tartu, DataDOI, doi:10.15155/re-72 (Estonian Landboard, 2017).

Author contributions.

A. Kmoch designed the experiments and code. A. Kmoch, E. Uuemaa, A. Astover, A. Kanal validated soil types and texture values. E. Uuemaa completed terrain analysis and statistics. A. Kull, A. Helm, M. Pärtel, I. Ostonen cleaned and organized the input SOC datasets. H. Virro developed the SOC RF machine-learning code and experiment. A. Kmoch and E. Uuemaa drafted the manuscript and created the figures, and all authors contributed to the paper writing.

Competing interests.

The authors declare that they have no conflict of interest.

Special issue statement.

This article is part of the special issue “Linking landscape organisation and hydrological functioning: from hypotheses and observations to concepts, models and understanding”. It is not associated with a conference.

Acknowledgements.

Financial support.

This research has been supported by the Marie Skłodowska-Curie Actions individual fellowships under the Horizon 2020 Programme grant agreement number 795625, the Mobilitas Pluss Postdoctoral Researcher Grant numbers MOBJD233 and PRG352 of the Estonian Research Council (ETAG), the European Regional Development Fund (Centre of Excellence EcolChange), and by the Estonian Environmental Investment Centre.

References

Abdelbaki, A. M.: Evaluation of pedotransfer functions for predicting soil bulk density for U.S. soils, Ain Shams Eng. J., 9(4), 1611–1619, doi:10.1016/j.asej.2016.12.002, 2018.

Adams, W. A.: The Effect of Organic Matter on the bulk and true Densities of some Uncultivated Podzolic Soils, J. Soil Sci., 24(1), 10–17, doi:10.1111/j.1365-2389.1973.tb00737.x, 1973.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrol. Sci. Bull., 24(1), 43–69, doi:10.1080/02626667909491834, 1979.

Breiman, L.: Random Forests, Mach. Learn., 45(1), 5–32, doi:10.1023/A:1010933404324, 2001.

Calhoun, T. E., Ellermäe, O., Kõlli, R., Lemetti, I., Penu, P. and Smith, C. W.: Benchmark Soils of Estonia Researched thru Baltic –American Collaboration. Problems of Estonian Soil Classification, Trans. Est. Agric. Univ., 198, 76–114, 1998.

Caruana, R. and Niculescu-Mizil, A.: An Empirical Comparison of Supervised Learning Algorithms, in Proceedings of the 23rd International Conference on Machine Learning, pp. 161–168, ACM, New York, NY, USA., 2006.

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(7), 1991–2007, doi:10.5194/gmd-8-1991-2015, 2015.

Dipak, S. and Abhijit, H.: Physical and Chemical Methods in Soil Analysis -, New Age International Ltd., New Delhi., 2005.

Ditzler, C., Scheffe, K. and Monger, H. C.: Soil survey manual. USDA Handbook 18, Soil Science Division. Government

Printing Office, Washington, D.C., 2017.

Estonian Landboard: Soilmap of Estonia - Mullastiku kaart, , doi:http://dx.doi.org/10.15155/re-72, 2017.

Eswaran, H., Van Den Berg, E. and Reich, P.: Organic Carbon in Soils of the World, Soil Sci. Soc. Am. J., 57(1), 192, doi:10.2136/sssaj1993.03615995005700010034x, 1993.

FAO: World reference base for soil resources 2014 International soil classification system., 2015.

Fischer, G., Nachtergaele, F., Prieler, S., van Velthuizen, H. T., Verelst, L. and Wiberg, D.: Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008), in IIASA, Laxenburg, Austria and FAO, Rome, Italy., 2008.

Gafni, A., Malterer, T., Verry, E., Nichols, D., Boelter, D. and Päivänen, J.: Physical Properties of Organic Soils, in Peatland Biogeochemistry and Watershed Hydrology at the Marcell Experimental Forest, edited by R. Kolka, S. Sebestyen, E. S. Verry, and K. Brooks, pp. 135–176, CRC Press, Boca Raton, FL., 2011.

Gunarathna, M. H. J. P., Sakai, K., Nakandakari, T., Momii, K. and Kumari, M. K. N.: Machine learning approaches to develop pedotransfer functions for tropical Sri Lankan soils, Water (Switzerland), doi:10.3390/w11091940, 2019.

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.

Hiederer, R., Michéli, E. and Durrant, T.: Evaluation of BioSoil DemonstrationProject, Ispra., 2011.

Kachinsky, N.: Fizika potchv. Soil physics. In Russian, Vol. 1. in., Moscow University Press, Moscow., 1965.

Kask, R.: On the English Equivalents of the Estonian Terms for the Textural Classes of Estonian Soils, J. Agric. Sci., 14, 93–96 [online] Available from: http://agrt.emu.ee/pdf/proceedings/toim_2001_14_kaskr.pdf, 2001.

Kauer, K., Astover, A., Viiralt, R., Raave, H. and Kätterer, T.: Evolution of soil organic carbon in a carbonaceous glacial till as an effect of crop and fertility management over 50 years in a field experiment, Agric. Ecosyst. Environ., 283, 106562, doi:10.1016/j.agee.2019.06.001, 2019.

Keesstra, S., Mol, G., de Leeuw, J., Okx, J., Molenaar, C., de Cleen, M. and Visser, S.: Soil-related sustainable development goals: Four concepts to make land degradation neutrality and restoration work, Land, doi:10.3390/land7040133, 2018.

Keesstra, S. D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J. N., Pachepsky, Y., Van Der Putten, W. H., Bardgett, R. D., Moolenaar, S., Mol, G., Jansen, B. and Fresco, L. O.: The significance of soils and soil science towards realization of the United Nations sustainable development goals, SOIL, doi:10.5194/soil-2-111-2016, 2016.

Kmoch, A., Kanal, A., Astover, A., Kull, A., Virro, H., Helm, A., Pärtel, M., Ostonen, I. and Uuemaa, E.: EstSoil-EH v1.0: An eco-hydrological modelling parameters dataset derived from the Soil Map of Estonia (data deposit), , doi:10.5281/zenodo.3473289, 2019a.

Kmoch, A., Virro, H. and Uuemaa, E.: EstSoil-EH v1.0 software supplement release for deposit, , doi:10.5281/ZENODO.3473210, 2019b.

Kõlli, R., Ellermäe, O., Köster, T., Lemetti, I., Asi, E. and Kauer, K.: Stocks of organic carbon in Estonian soils, Est. J. Earth Sci., 58(2), 95, doi:10.3176/earth.2009.2.01, 2009.

Kriiska, K., Frey, J., Asi, E., Kabral, N., Uri, V., Aosaar, J., Varik, M., Napa, Ü., Apuhtin, V., Timmusk, T. and Ostonen, I.: Variation in annual carbon fluxes affecting the SOC pool in hemiboreal coniferous forests in Estonia, For. Ecol. Manage., 433, 419–430, doi:10.1016/j.foreco.2018.11.026, 2019.

Laas, A. and Kull, A.: Sustainable Planning and Development, edited by A. G. K. E. Beriatos, C.A. Brebbia, H. Coccossis, Boston: Wessex Institute of Techonology Press, Southampton., 2003.

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.

Michielsen, A., Kalantari, Z., Lyon, S. W. and Liljegren, E.: Predicting and communicating flood risk of transport infrastructure based on watershed characteristics, J. Environ. Manage., doi:10.1016/j.jenvman.2016.07.051, 2016.

Mokarram, M., Roshan, G. and Negahban, S.: Landform classification using topography position index (case study: salt dome of Korsia-Darab plain, Iran), Model. Earth Syst. Environ., doi:10.1007/s40808-015-0055-9, 2015.

Moore, I. D., Grayson, R. B. and Ladson, A. R.: Digital terrain modelling: A review of hydrological, geomorphological, and biological applications, Hydrol. Process., doi:10.1002/hyp.3360050103, 1991.

Noreika, N., Helm, A., Öpik, M., Jairus, T., Vasar, M., Reier, Ü., Kook, E., Riibak, K., Kasari, L., Tullus, H., Tullus, T., Lutter, R., Oja, E., Saag, A., Randlane, T. and Pärtel, M.: Forest biomass, soil and biodiversity relationships originate from biogeographic affinity and direct ecological effects, Oikos, oik.06693, doi:10.1111/oik.06693, 2019.

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, doi:10.5194/soil-4-1-2018, 2018.

Prévost, M.: Predicting Soil Properties from Organic Matter Content following Mechanical Site Preparation of Forest Soils, Soil Sci. Soc. Am. J., 68(3), 943, doi:10.2136/sssaj2004.9430, 2004.

Putku, E.: Prediction models of soil organic carbon and bulk density of arable mineral soils, Estonian University of Life Sciences., 2016.

Reintam, L., Kull, A., Palang, H. and Rooma, I.: Large-Scale Soil Maps and a Supplementary Database for Land Use Planning in Estonia, J. Plant Nutr. Soil Sci. Fur Pflanzenernahrung Und Bodenkd., 166(2), 225−231, 2003.

Reintam, L., Rooma, I., Kull, A. and Kõlli, R.: Soil information and its application in Estonia, in Research report, vol. 9, edited by European Soil Bureau, pp. 121–132., 2005.

Suuster, E., Ritz, C., Roostalu, H., Reintam, E., Kõlli, R. and Astover, A.: Soil bulk density pedotransfer functions of the humus horizon in arable soils, Geoderma, 163(1–2), 74–82, doi:10.1016/j.geoderma.2011.04.005, 2011.

Tarnocai, C., Canadell, J. G., Schuur, E. A. G., Kuhry, P., Mazhitova, G. and Zimov, S.: Soil organic carbon pools in the northern circumpolar permafrost region, Global Biogeochem. Cycles, 23(2), doi:10.1029/2008GB003327, 2009.

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

Uuemaa, E., Hughes, A. O. and Tanner, C. C.: Identifying feasible locations for wetland creation or restoration in catchments by suitability modelling using light detection and ranging (LiDAR) Digital Elevation Model (DEM), , 10(4), doi:10.3390/w10040464, 2018.

Vitharana, U. W. A., Mishra, U., Jastrow, J. D., Matamala, R. and Fan, Z.: Observational needs for estimating Alaskan soil carbon stocks under current and future climate, J. Geophys. Res. Biogeosciences, doi:10.1002/2016JG003421, 2017.

Yigini, Y. and Panagos, P.: Assessment of soil organic carbon stocks under future climate and land cover changes in Europe, Sci. Total Environ., 557–558, 838–850, doi:10.1016/J.SCITOTENV.2016.03.085, 2016.

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, 39–53, doi:10.1016/j.jhydrol.2017.01.004, 2017.