International Soil and Water Conservation Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Soil and Water Conservation Research

journal homepage: www.elsevier.com/locate/iswcr

Original Research Article

Global mapping of volumetric water retention at 100, 330 and 15 000 cm suction using the WoSIS database

Maria Eliza Turek [a, b, *, 1], Laura Poggio [a], Niels H. Batjes [a], Robson André Armindo [d], Quirijn de Jong van Lier [e], Luis de Sousa [a], Gerard B.M. Heuvelink [a, c]

[a] ISRIC - World Soil Information, Wageningen, the Netherlands
[b] Graduate Program in Environmental Engineering, Federal University of Paraná, Curitiba, PR, Brazil
[c] Soil Geography and Landscape Group, Wageningen University, Wageningen, the Netherlands
[d] Department of Physics, Federal University of Lavras, MG, Brazil
[e] CENA - University of São Paulo, Piracicaba, SP, Brazil

ARTICLE INFO

Article history:
Received 14 February 2022
Received in revised form
6 July 2022
Accepted 3 August 2022
Available online xxx

Keywords:
Digital soil mapping
Soil hydraulic properties
Pedometrics
SoilGrids

ABSTRACT

Present global maps of soil water retention (SWR) are mostly derived from pedotransfer functions (PTFs) applied to maps of other basic soil properties. As an alternative, 'point-based' mapping of soil water content can improve global soil data availability and quality. We developed point-based global maps with estimated uncertainty of the volumetric SWR at 100, 330 and 15 000 cm suction using measured SWR data extracted from the WoSIS Soil Profile Database together with data estimated by a random forest PTF (PTF-RF). The point data was combined with around 200 environmental covariates describing vegetation, terrain morphology, climate, geology, and hydrology using DSM. In total, we used 7292, 33 192 and 42 016 SWR point observations at 100, 330 and 15 000 cm, respectively, and complemented the dataset with 436 108 estimated values at each suction. Tenfold cross-validation yielded a Root Mean Square Error (RMSE) of 6.380, 7.112 and 6.485 10^{-2}cm^{3}cm^{-3}, and a Model Efficiency Coefficient (MEC) of 0.430, 0.386, and 0.471, respectively, for 100, 330 and 15 000 cm. The results were also compared to three published global maps of SWR to evaluate differences between point-based and map-based mapping approaches. Point-based mapping performed better than the three map-based mapping approaches for 330 and 15 000 cm, while for 100 cm results were similar, possibly due to the limited number of SWR observations for 100 cm. Major sources or uncertainty identified included the geographical clustering of the data and the limitation of the covariates to represent the naturally high variation of SWR.

1. Introduction

Soil water retention controls multiple processes related to mass and energy cycles in the soil-plant-atmosphere system. It impacts the exchange of gases (de Jong van Lier et al., 2018), including trace gases such as carbon dioxide. It influences soil water availability, which is considered an important ecosystem service, essential to nutrient cycling and primary production (Dobriyal et al., 2012), provisioning of food, feed, fiber and fuel as well as regulation of climate, gas exchange, and water flows, control of erosion and flooding and water purification (Adhikari & Hartemink, 2016). Modelling water retention at specific suctions is helpful for determining the available water capacity (AWC), defined as the difference between upper and lower limits of water retention, on a volumetric basis. AWC is an important parameter in bucket-type models, used in crop (e.g., DSSAT (Hoogenboom et al., 1999), AquaCrop (Raes et al., 2016, pp. 1–19)) and ecological (e.g., Laio et al. (2001); Pumo et al. (2008)) applications. Quantifying the spatial variation of AWC is important for planning and risk

* Corresponding author. ISRIC - World Soil Information, Wageningen, the Netherlands.
E-mail addresses: mariaeliza.turek@agroscope.admin.ch (M.E. Turek), laura.poggio@wur.nl (L. Poggio), niels.batjes@isric.org (N.H. Batjes), robson.armindo@ufla.br (R.A. Armindo), qdjvlier@usp.br (Q. de Jong van Lier), luis.desousa@isric.org (L. de Sousa), gerard.heuvelink@wur.nl (G.B.M. Heuvelink).
[1] Present address: Agroscope, Division of Agroecology and Environment, Group of Climate and Agriculture, Zürich, Switzerland.

https://doi.org/10.1016/j.iswcr.2022.08.001

mitigation purposes (Poggio et al., 2010). Water retention as a function of soil water suction is a key component in both Richards equation-based and bucket-type water balance models that represent the processes in vadose zone hydrology (Vereecken et al., 2008), eco-hydrological (Porporato et al., 2015), and agro-hydrological (Eitzinger et al., 2004) evaluations. There is also a marked demand for global maps of soil hydraulic properties in land surface models (LSMs), which are a key component in Earth System modelling (Dai, Shangguan, et al., 2019), with emphasis on the soil water retention data.

At the point scale, the techniques for measuring soil water retention have advanced, but direct determination is still expensive, time-consuming, and impractical for large-scale applications (Vereecken et al., 2010). As an alternative to direct measurements, pedotransfer functions (PTFs) have been widely used at different scales (Van Looy et al., 2017). As these are generally mere empirical-statistical relationships, their accuracy outside the development database range is essentially unknown (Vereecken et al., 2016). Due to this 'limited portability', their widespread implementation remains a challenge. According to Dai, Xin, et al. (2019), application of a single-PTF to globally predict soil hydraulic properties leads to biases, underestimation of uncertainties, and overconfidence in model performances of predictive applications.

Present global maps of water retention generally show the parameters of a soil water retention (SWR) function that were obtained by PTFs, implying that the uncertainty of the resulting maps depends on the accuracy of the PTFs (Padarian et al., 2014) as well as the underpinning soil data. Table 1 shows examples of studies that present global estimates of soil water retention based on PTFs. A similar overview for studies at the regional and national scale can be found in Dai, Xin, et al. (2019).

An alternative for the use of PTFs for mapping is to collect sufficient direct observations of soil water retention and use these for producing interpolated maps using digital soil mapping (DSM) techniques (Leenaars et al., 2018). However, considering the usual lack of measured data and the high spatial variability, this direct method is more common for 'smaller' areas. Padarian et al. (2014), for example, applied DSM to measured values of water retention at field capacity (FC) and permanent wilting point (PWP) to produce maps of soil AWC for Australia's wheat belt. Similarly, Vasques et al. (2016) presented maps of soil water retention at 100 and 15 000 cm suction in a tropical dry forest in Brazil, while Dharumarajan et al. (2020) developed maps of FC and PWP for the Northern Karnataka Plateau, India, with uncertainty estimates derived from a quantile regression random forest model.

Considering that both point-based and map-based approaches to obtain maps of soil water retention are influenced by PTFs, in this study, we aimed to test if the point-based mapping of soil water retention is a suitable approach at the global scale. As a point-based approach, we consider the interpolation of soil water retention data, both measured and estimated, using DSM approaches. Alternatively, in the map-based approach soil maps with basic properties are used as an input to published PTFs.

Using the point-based approach, we created maps based on data from the WoSIS global soil database (Batjes et al., 2020), a set of covariates, and machine learning techniques building on procedures developed for the SoilGrids project (Poggio et al., 2021). Subsequently, the generated maps were compared with three published PTF-derived global SWR maps, using point accuracy metrics and qualitative comparison of spatial patterns.

We considered water retention at three commonly measured water suctions (100, 330 and 15 000 cm) usually adopted as thresholds for calculating the AWC, following the United States Department of Agriculture (USDA) (Soil Survey Staff, 2014) conventions.

2. Materials and methods

Production of maps and their evaluation was performed following four main steps: 1. Screening and selection of available water retention data in WoSIS, including the development of a pedotransfer function (PTF) to estimate soil bulk density as needed to convert gravimetric into volumetric water retention; 2. Development of a Random Forest-based PTF (PTF-RF) to obtain volumetric water retention from other soil properties available in WoSIS to generate more data for the mapping; 3. Generate SWR maps with associated prediction uncertainty; 4. Compare the mapping results with other published products.

In all steps where we assess models by comparing predicted with observed data, we used common accuracy measures such as the root mean squared error (RMSE), mean error (ME) and the model efficiency coefficient (MEC, Janssen and Heuberger (1995)). MEC is defined as 1 minus the ratio between the error sum of squares and the total sum of squares, the fraction of the explained

Table 1
Summary of maps with global data of soil water retention.

Reference	Name	Base map	SWRC model	PTF
De Lannoy et al. (2014)	An updated treatment of soil texture and associated hydraulic properties in a global land modeling system	Harmonized World Soil Database version 1.21 (HWSD1.21) and the State Soil Geographic (STATSGO2)	Campbell (1974)	Wösten et al. (2001)
Montzka et al. (2017)	Global soil hydraulic properties and sub-grid variability of soil water retention and hydraulic conductivity curves	SoilGrids 1-km	Van Genuchten (1980) with the Mualem (1976) parameter restriction (VGM)	Rosetta
Zhang et al. (2018)	High-Resolution Global Map of Soil Hydraulic Properties Produced by a Hierarchical Parameterization of a Physically Based Water Retention Model	SoilGrids 1-km	θ(330 cm) and θ(15 000 cm)	Zhang et al. (2018)
Dai et al. (2019b)	Global High Resolution Data Set of Soil Hydraulic and Thermal Properties for Land Surface Modeling	Global Soil Dataset for Earth System Models (GSDE)	Campbell (1974)	Ensamble
Han et al. (2019)	Global High-Resolution Soil Profile Database for Crop Modeling Applications	SoilGrids 1-km	VGM θ(330 cm) and θ(15 000 cm)	Ensamble Saxton and Rawls (2006)
Simons et al. (2020)	HiHydroSoil v2.0 - High Resolution Soil Maps of Global Hydraulic Properties	SoilGrids250m-2.0	VGM	Tóth et al. (2015)
Zhang et al. (2020)	Development of Hierarchical Ensemble Model and Estimates of Soil Water Retention with Global Coverage	OpenGeoHub	θ(330 cm) and θ(15 000 cm)	Ensamble
Reynolds et al. (2000)	Estimating soil water-holding capacities by linking the Food and Agriculture Organization soil map of the world with global pedon databases and continuous pedotransfer functions	FAO soil map of the world	Available water holding capacity	Saxton et al. (1986)

variance based on the 1:1 line of predicted versus observed. In hydrology it is also termed the Nash-Sutcliffe model efficiency (Nash & Sutcliffe, 1970). We predicted the SWR parameters at the centres of standard depth intervals following the specifications of GlobalSoilMap (Arrouays et al., 2014), namely 0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm and 100–200 cm. These point predictions were used as proxies of interval averages.

2.1. Data selection and harmonisation

The World Soil Information Service (WoSIS) (Batjes et al., 2020) provides a compilation of freely shared soil profile data that have been standardized for global applications, including soil water retention data. The soil water retention data in the source datasets submitted to WoSIS were either provided on a volumetric basis (i.e. pre-converted from gravimetric data) or more commonly on a gravimetric basis, whereas many applications require soil water retention expressed on a volumetric basis. Due to the shortage in volumetric water retention data, an approach to include the gravimetric water retention data was developed, with the conversion to volumetric content performed based on systematic rules and consistency checks, as described below.

Regarding the sampling type of data in gravimetric measurements, for the lower suctions (i.e. 100 cm and 330 cm), we considered measurements on undisturbed clod and core samples as soil structure is important at these low suctions. Alternatively, at 15 000 cm, where soil texture plays a greater role than soil structure, disturbed samples were mainly used.

Conversely, when the data were submitted on a volumetric basis (i.e. previously converted from gravimetric by data providers) it proved more difficult to filter out whether any disturbed samples were used for the lower suctions (i.e. 100 and 330 cm), based on the information provided with the source materials. Where specified, more than 97 percent of the previously converted volumetric data were reported for undisturbed samples. We assumed this to apply also to those sets for which no information on the conversion was provided, which is a simplification. Pragmatically, we assumed that any data for disturbed samples (at lower suctions) would largely be filtered out during the subsequent cleaning process (Fig. 2). It should be noted, however, that this type of methodological 'unknowns' will always occur in global data compilations as the metadata for the source data are seldom complete; this aspect will be reflected in the possible accuracy of the present global predictions as discussed in Section 3.

Based on the above, we selected soil profiles with a) water retention data already reported on a volumetric basis (θ, 10^{-2} cm^3 cm^{-3}) and b) data reported on a gravimetric (w, 10^{-2} g g^{-1}) basis at soil water suctions ($h > 0$ in the unsaturated soil) of 100, 330 and 15 000 cm. For case b), ultimately, soil bulk density using volume at 330 cm (ρ_{330}) (Table 2) were used to convert the gravimetric data to volumetric water retention. Observations were allocated to the interval that contained the midpoint depth of the measured soil layer. Fig. 1 shows the resulting distribution of the number of observations per standard depth interval. The number of observations is well distributed between the standard depth intervals, although there are fewer observations for the surface layer (0–5 cm), due to its shallowness.

Fig. 2 describes the major steps of the procedure for selecting, screening and converting the available data. For each suction, a plausible minimum (θ_{min}) and maximum value (θ_{max}) were defined according to thresholds commonly reported in the literature (Minasny et al., 2004; Nemes et al., 2001; Twarakavi et al., 2009). For the lower suctions (100 and 330 cm), θ_{max} was set at 80 10^{-2} cm^3 cm^{-3} and θ_{min} at 1 10^{-2} cm^3 cm^{-3}. At higher suction (15 000 cm), θ_{min} was set at zero and θ_{max} at 60 10^{-2} cm^3 cm^{-3}. These thresholds were used to check the volumetric water retention data, both as already converted as well as those derived from the gravimetric water retention data (Section 2.1.1). Considering case a), the procedure led to the exclusion of thirty layers (0.17%) with pre-converted volumetric data at 15 000 cm suction. For 100 and 330 cm, no layers needed to be excluded.

2.1.1. Converting gravimetric into volumetric water retention

As indicated, part of the source data were reported on a gravimetric basis. The conversion from gravimetric (w, 10^{-2} g g^{-1}) into volumetric (θ, 10^{-2} cm^3 cm^{-3}) water retention was performed using the soil bulk density (ρ_{330}), defined as the mass of dry soil divided by the sample volume at $h = 330$ cm, following USDA (Soil Survey Staff, 2014) standards. According to Nemes et al. (2010), ρ_{330} is a better measure to represent field conditions than the bulk density determined after a sample has been exposed to extreme shrinkage conditions in an oven at 105 °C.

The availability of water retention data per soil layer varied greatly between soil profiles. For some layers (or horizons), both gravimetric water retention and ρ_{330} measured data were available, for others gravimetric water retention and only ρ_b measured data were found, and yet others contained no information at all about soil bulk density. When ρ_{330} measured was available, possibly 'suspicious' values were eliminated when they did not meet the $\rho_{330} \leq c \cdot \rho_b$ criterion. Factor $c = 1.15$ is an arbitrary tolerance level associated with possible measurement errors for ρ_{330} and ρ_b themselves (see S1 in Supplementary materials).

Further, we checked if the calculated volumetric water retention was within common limits $\theta_{min} \leq \theta \leq \theta_{max}$ and not higher than total porosity φ, which represents the maximum volumetric fraction of pores for the given bulk density. Here it was calculated as $\varphi = 1 - \rho_{330}/\rho_p$, in which ρ_p is the particle density taken as 2.65 g cm^{-3} (Blake, 2008, pp. 504–505). From the original sets for 100, 330 and 15 000 cm suction with, respectively, 3 522, 91 808 and 20 447 observations with measured ρ_{330} values, 33.8, 29.4 and 0.8% did not satisfy the conditions for 100, 330 and 15 000 cm, respectively. Assuming an error that could be derived from ρ_{330} measurements, these observations were merged with the data that had only ρ_b measured and where ρ_{330} was estimated with a linear regression pedotransfer function: $\rho_{330} = 0.2308 + 0.8253\rho_b$, with residual standard deviation (RSE) of 0.1254 g cm^{-3} and model efficiency coefficient (MEC) of 0.7525. The details of the data used to this PTF development are presented in the Supplementary Materials (Section S1). In short, we considered only ρ_b as predictor variable similar to Heuscher et al. (2005), but unlike other studies that used more complex PTFs (de Souza et al., 2016; Nemes et al., 2010; Sequeira et al., 2014; Seybold et al., 2014). This was done because a more complex PTF would be applicable only to a small portion of the data.

Although we make use of a global PTF to estimate ρ_{330}, we expect that inconsistencies are likely to be filtered in the subsequent steps, as the volumetric water retention calculated from the predicted ρ_{330} was evaluated in the same way as those obtained from the measured ρ_{330} data. From the original 3 449 (100 cm), 16 997 (330 cm) and 12 308 (15 000 cm) values with ρ_{330} estimated, 60.1, 36.0, and 92.3% respectively met all consistency rules; these values were kept for further analysis. Data that did not fulfill the consistency rules and those lacking information for ρ_{330} and ρ_b were discarded. Subsequently, the 'already converted' measured volumetric water retention data were merged with those derived from the gravimetric data. The merged data was evaluated for inconsistencies using the rules described earlier: 252 pairs of data were removed because θ(100 cm) < θ(330 cm), 519 pairs because θ(100 cm) < θ(15 000 cm) and 1107 because θ(330 cm) < θ(15 000 cm).

Table 2
Summary statistics for water retention data in WoSIS, at suctions of 100, 330 and 15 000 cm, reported on a volumetric (θ) and gravimetric (w) basis; soil bulk density measured at 330 cm (ρ_{330}) and at oven-dry conditions (ρ_b).

	N	Mean	SD	Min	$q_{0.25}$	Median	$q_{0.75}$	Max	Skewness
w (g 100g^{-1})									
100 cm	12552	24.70	13.00	1.00	15.00	23.70	32.30	80.00	0.80
330 cm	94923	25.10	11.50	1.00	17.50	23.90	30.90	80.00	1.00
15 000 cm	182681	12.60	8.90	1.00	6.70	10.90	16.20	80.00	2.40
θ (cm 100 cm^{-1})									
100 cm	5217	35.80	13.30	1.00	27.00	35.50	43.40	80.00	0.30
330 cm	17569	26.20	14.30	1.00	15.00	25.10	35.00	80.00	0.60
15 000 cm	17578	16.10	11.00	0.00	8.00	14.00	22.70	71.00	1.00
ρ (g cm^{-3})									
ρ_{330}	2951	1.51	0.24	0.05	1.41	1.53	1.65	2.41	−1.26
ρ_b	123538	1.42	0.37	0.01	1.28	1.50	1.66	2.63	−1.28

N: Number of available layers, SD: standard deviation, Min: minimum value, Max: maximum value.

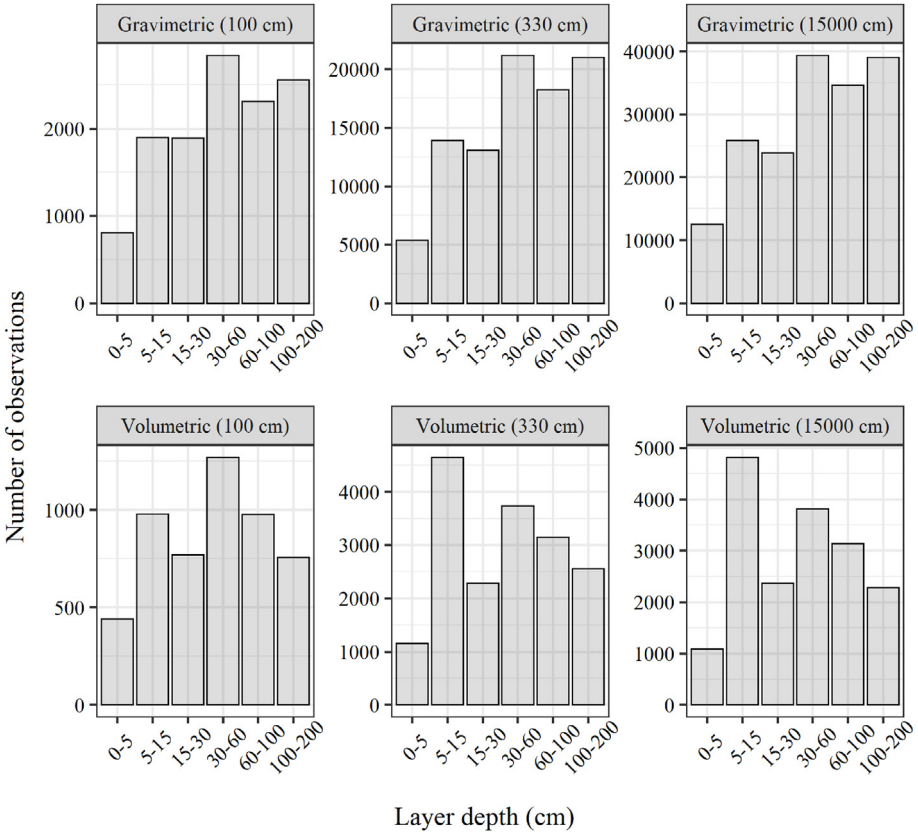

Fig. 1. Distribution of the volumetric and gravimetric water retention data in WoSIS for the standard depth intervals considering the suctions of 100, 330 and 15 000 cm.

2.2. Pedotransfer function to estimate soil water retention at specific points

The merged dataset built with gravimetric and volumetric data from WoSIS was evaluated in a prior attempt to generate global maps using only measured data, which lead to inconsistencies due to the sparse available data, as discussed in the Supplementary Materials (Section S2). To further increase the dataset for mapping, a PTF-RF (Wright & Ziegler, 2017) for predicting volumetric soil water retention from basic soil properties was calibrated using global soil data from the cleaned dataset. The PTF-RF used clay content, silt content, soil organic carbon content and pH_{H2O} from WoSIS as predictor variables. This combination was selected based on the performance of the PTF-RF, the availability of data for developing the PTF-RF and the number of observations that could be added to the dataset using this PTF-RF. The tested combinations were evaluated using ten-fold cross-validation. The folds were built considering soil observations spatially stratified in the geodetic domain to guarantee a balanced spatial distribution within each fold (Poggio et al., 2021). More details about the selection of the predictors and PTF-RF performance are given in the Supplementary Materials (Section S2).

2.3. Mapping

The geo-referenced dataset (see column 1 Table 5), with the reported and PTF-RF-predicted volumetric soil water retention data, was used for mapping of volumetric water retention,

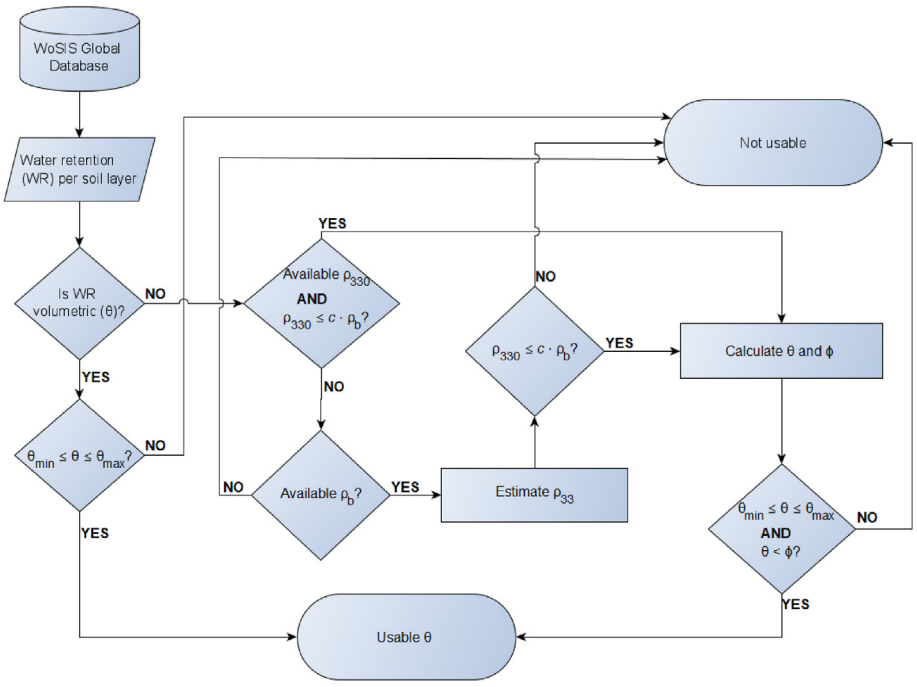

Fig. 2. Flowchart for developing a consistent set of soil water retention data on a volumetric basis (θ). ρ_{330}:soil bulk density with volume at 330 cm, ρ_b: dry soil bulk density, φ: soil total porosity. θ_{min} and θ_{max} are predefined thresholds for θ depending on the suction. c was assumed equal to 1.15.

following the procedures used in SoilGrids 2.0 (Poggio et al., 2021). A machine learning approach using Random Forest (Breiman, 2001) was applied. Accordingly, around 200 covariates were prepared as candidate predictors based on their likely influence on soil formation. From this initial pool, covariates were selected to develop a parsimonious and computationally efficient model, decrease the risk of over-fitting and reduce bias in variable importance assessment. Two steps were used: 1. de-correlation: only covariates with a pairwise correlation coefficient $\leq$0.85 were selected; 2. recursive feature elimination: the importance of the covariates was assessed in recursive loops and less important covariates were removed by minimizing the RMSE.

Model tuning of the Random Forest model was performed considering different combinations of hyper-parameters, in particular the number of decision trees (*ntree*) and the number of covariates used in the tree splits (*mtry*). This was done using ten-fold cross-validation, with observations in each fold spatially stratified in the geodetic domain. The results of each combination were evaluated according to the root mean squared error (RMSE) and model efficiency coefficient (MEC). The final model was fitted with all available water retention data, the selected covariates and the optimized hyper-parameters. Observation depth was included as a covariate, and calculated as the midpoint of the sampled soil layer. The final prediction models were generated using the `ranger` package as computationally optimized implementation of Random Forest (Wright & Ziegler, 2017). The option quantreg was used to build quantile regression forests. This yields a cumulative probability distribution of soil water retention at each location and depth, thus also quantifying prediction uncertainty (Poggio et al., 2021).

Maps of the mean, median (0.50 quantile, $q_{0.50}$), 0.05 quantile ($q_{0.05}$), and 0.95 quantile ($q_{0.95}$) were produced at 250 m resolution for all six standard depths. The uncertainty of the maps was quantified by the prediction interval ratio (PIR), defined as the ratio of the 90% prediction interval width and the median:

$$\mathrm{PIR} = \frac{q_{0.95} - q_{0.05}}{q_{0.50}} \quad (1)$$

2.4. Comparison of point-based with map-based derived maps and map evaluation

The above point-based obtained maps, hereafter referred to as 'SoilGrids', were compared with three existing map-based derived maps selected from Table 1. This comparison was performed in two ways: 1) map-to-map comparison, using an equally distributed grid of points and, 2) comparison with point measurements, where we evaluated the accuracy of the maps in predicting observed data from the WoSIS database.

Here we use the term map-based when maps are created using indirect methods such as pedotransfer functions (PTFs), applied to maps of basic soil properties (Dai, Xin, et al., 2019; Szabó et al., 2019). The map-based derived maps will be referred to as: 'Map1' (Dai, Xin, et al., 2019), 'Map2' (Montzka et al., 2017) and 'Map3' (Simons et al., 2020). The main characteristics of the four maps are presented in Table 3.

Map1 is part of a global high-resolution dataset of soil hydraulic and thermal parameters. It was built using an ensemble of PTFs applied to maps of basic soil properties from the Global Soil dataset for Earth System Models (GSDE) (Shangguan et al., 2014). We used the mapped parameters of the soil water retention function described by the Van Genuchten (1980) model with the Mualem (1976) parameter restriction (VGM) to calculate the water retention at 100, 330 and 15 000 cm at the evaluated points. The depth intervals were the same as those used in this study.

Map2 is part of a global dataset of soil hydraulic properties and sub-grid variability of soil water retention and hydraulic conductivity. The maps were built to demonstrate a method to scale hydraulic parameters to individual model grids and provide a global

dataset using the Rosetta PTF (Schaap et al., 2001), applied to the 'SoilGrids-1km' (Hengl et al., 2014) dataset. We used the mapped VGM parameters to calculate the water retention at 100, 330 and 15 000 cm at the evaluated points. Map2 was originally presented at seven depths (0, 5, 15, 30, 60, 100, and 200 cm); we took the mean of the predictions at the top and bottom of each depth interval to obtain predictions at the standard depth intervals considered in this study.

Map3 is a global dataset of soil hydraulic properties based on the application of parametric PTFs developed for Europe (Tóth et al., 2015) to the most recent release of SoilGrids (Poggio et al., 2021). The maps with the VGM parameters were used to calculate the volumetric water retention at 100, 330 and 15 000 cm.

The map comparison considered a grid created with an Icosahedral Snyder Equal-Area Grid (ISEAG) of resolution 10, resulting in 590 492 strata (i.e. hexagonal cells), each with an area of about 900 km^2. This grid was chosen to avoid distortions and to guarantee that all grid cells have the same size. This grid was similar to the one presented in Fig. 4. The centroid of each hexagon was derived and used to extract the corresponding values from the considered maps. All points were reprojected to match the map projection, to reduce artifacts inherent with re-projection of raster layers. The resulting point datasets were compared by density plots, spatial distribution plots, and scatter density plots of map-based maps against the SoilGrids map.

Besides map inter-comparison, the accuracy of the four products was evaluated using observed data points from the WoSIS database. For the SoilGrids maps, the evaluation was performed in 10-fold cross-validation mode, meaning that predictions were derived from a model that was calibrated on 9 folds, thus always excluding the fold that contains the validation data. This was done to avoid overoptimistic validation results for SoilGrids. For the other products, the validation points were reprojected to match the map projection. Scatter density plots were made and prediction performance was evaluated using the RMSE, ME and MEC.

As the available database presents a higher density of observations in the USA than in the rest of the world, the performance of the four maps according to RMSE was also computed separately considering data from 'USA only' and from the 'rest of the world'. Both the inter-comparison between maps and the evaluation of the map accuracy were applied to the six GlobalSoilMap standard depths.

2.5. Computational environment

The screening, selection and merging of the WoSIS datasets as well as other dataset manipulations were performed using `R` (R Core Team, 2020), in particular functions available in the `tidyverse` set of packages (Wickham et al., 2019). The PTF-RF to estimate soil water retention was built using the `ranger` (Wright & Ziegler, 2017) package. Subdivision of the locations in the folds, while maintaining the spatial distribution, was performed using the `caret` (Kuhn, 2021) package. Manipulations of raster type data were performed using the `terra` (Hijmans, 2021b) and `raster` (Hijmans, 2021a) packages. The grid used for map comparison was created with the `dggridR` package (Barnes et al., 2020). Evaluation of the cross-validation datasets and the map comparisons were performed on a point-basis, using the `sf` (Pebesma, 2018) and `rgdal` (Bivand et al., 2021) packages for the necessary transformations between map projections and for data storage as geopackages. Plotting of the final maps was performed with the `tmap` (Tennekes, 2018) and `tmaptools` (Tennekes, 2021) packages.

The mapping was performed according to the SoilGrids 2.0 workflow (Poggio et al., 2021) with a dynamic geographic tiling system and a parallelisation scheme described in de Sousa et al. (2020). The code is available under the GPL3 license at the SoilGrids git repository.

Table 4
Cross-validation statistics for the PTF-RF developed to estimate soil water retention at layers lacking observed data. Values of RMSE and ME are in 10^{-2} cm^3 cm^{-3}, MEC is dimensionless.

	N (PTF-RF calibration)	RMSE	ME	MEC
θ (100 cm)	7 292	8.6	−0.1	0.525
θ (330 cm)	33 192	7.6	−0.1	0.567
θ (15 000 cm)	42 016	5.4	−0.1	0.681

RMSE: root mean error, ME: mean error, MEC: model efficiency coefficient.

3. Results and discussion

3.1. Merged datasets used for mapping soil water retention

Table 4 presents summary statistics for the PTF-RF derived on the WoSIS dataset to estimate water retention at 100, 330 and 15 000 cm suction from basic soil properties (Section 2.2). Bias was not observed, explaining between 52 and 69 per cent of the variation, and that predictions were best for 15 000 cm and worst for 100 cm. Table 5 presents summary statistics for the input data from the observed dataset, as derived from the original data described in Table 2 and application of the screening and merging procedure (Fig. 2). The number of layers with observed data for 100 cm were about 20% of that for layers with data for 330 and 15 000 cm suction. Table 5 also presents summary statistics for the estimated dataset obtained using the PTF-RF described in Section 2.2, for which the number of layers were always 436 108, irrespective of the suction under consideration, and corresponding to the number of available observations with the properties used in the PTF-RF. Overall, the PTF-RF-derived data were slightly more skewed than the observed data and the inter-quartile ranges were somewhat smaller, which may be due to the smoothing effects of PTFs.

Density plots and density scatter plots comparing the three suctions for the observed and estimated data (Fig. 3) showed a similar distribution between observed and estimated soil water retention data, despite the slightly different summary statistics (Table 5). For the observed data, Fig. 3 confirmed that θ(15 000 cm) $<$ θ(330 cm) $<$ θ(100 cm).

The geographical distribution of the observed and estimated volumetric water retention at 330 cm is shown in Fig. 4. Similar figures for the other two suctions are presented in the Supplementary Materials (Section S3). Most of the observed locations were in the USA with fewer observations for South America and

Table 3
Characteristics of global soil water retention maps used for the comparison. More details about the compared map-based products can be found in Table 1.

Name	Reference	Depth (cm)	Spatial resolution
Map1	Dai et al. (2019b)	0-5, 5−15, 15−30, 30−60, 60−100, 100-200	30"
Map2	Montzka et al. (2017)	0, 5, 15, 30, 60, 100, 200	0.25°
Map3	Simons et al. (2020)	0-5, 5−15, 15−30, 30−60, 60−100, 100-200	250 m
SoilGrids	this work	0-5, 5−15, 15−30, 30−60, 60−100, 100-200	250 m

Table 5
Summary statistics for volumetric water retention represented in the screened, merged dataset (observed) and in the dataset estimated by PTF-RF (estimated). Volumetric water retention (θ) in 10^{-2} cm^3 cm^{-3}.

	N	Mean	SD	Min	$q_{0.25}$	Median	$q_{0.75}$	Max	Skewness
Water content from observed data									
θ (100 cm)	7 292	33.8	12.5	2.7	25.4	34.1	41.5	79.1	0.24
θ (330 cm)	33 192	28.4	11.6	1.0	20.5	29.6	36.3	80.0	−0.04
θ (15 000 cm)	42 016	17.0	9.6	0.0	9.6	15.6	23.0	59.9	0.66
Water content from estimated data									
θ (100 cm)	436 108	35.3	8.9	3.6	30.8	37.6	41.1	76.0	−0.84
θ (330 cm)	436 108	29.8	9.8	1.3	23.9	32.2	37.0	73.0	−0.73
θ (15 000 cm)	436 108	16.4	8.9	0.0	9.5	15.4	22.3	52.8	0.46

N: Number of available layers, SD: standard deviation, Min: minimum value, Max: maximum value.

Asia, reflecting the regionally uneven distribution of the shared observed data in WoSIS. If a certain data-poor area has environmental conditions similar to a data-rich area, then spatial prediction for that data-poor area will benefit from the calibration data available for the data-rich area. In other situations, predictions in data-poor areas may have low accuracy because of extrapolations in feature space (Meyer & Pebesma, 2021). In such cases, the modelling and mapping would likely benefit from adding the PTF estimates of soil water retention, which are more uniformly distributed across the globe, as performed here.

3.2. Cross-validation and mapping of soil water retention

For the mapping itself, we used the volumetric water retention data derived according to Fig. 2, the so-called 'observed' data, combined with the set of PTF-RF-derived or 'estimated' data (Section 2.2). Summary statistics from the cross-validation are presented in Table 6.

The RMSE obtained for the estimated dataset was lower than for the observed dataset, which appears to be the result of the use of the PTF-RF, that tends to smooth the effects of natural soil heterogeneity. This trend was not observed for MEC because it scales the error variance of the data, which is smaller for the estimated dataset. The ME is small in all cases and hardly contributes to the RMSE. At 100 cm, even with the lower availability of observed data (Fig. S3), the SoilGrids model performed better than for other suctions, even though the water retention values were higher. This result shows the benefit provided by using the PTF-RF (Table 4) for improving the input dataset. Results of the cross-validation per depth interval, considering the complete dataset, are presented in the Supplementary Materials (Section S4). Scatter density plots based on only the observed dataset compared to the predictions from mapping are presented in Fig. 5. The figure confirmed large differences between SoilGrids predictions and independent observations, but show that there is no systematic prediction error.

The cross-validation showed the mapping procedure to be able to adequately explain the spatial variation of the soil properties. Despite this being a global analysis, the values reported for RMSE, ME and MEC were similar to those reported in studies at a regional scale. Dharumarajan et al. (2020) evaluated soil water retention in the Northern Karnataka Plateau, with RMSE depending on depth and soil suction varying from 4.71 to 7.38 10^{-2}cm^{3}cm^{-3} while Malone et al. (2020) predicted soil water retention across Australia's agricultural region with RMSE from 6.31 to 10.70 10^{-2}cm^{3}cm^{-3}, which are comparable accuracy to the results observed in this work. Metrics for the estimation of water retention at suctions closer to the permanent wilting point (15 000 cm) tended to be more accurate than those closer to saturation, with lower RMSE and ME

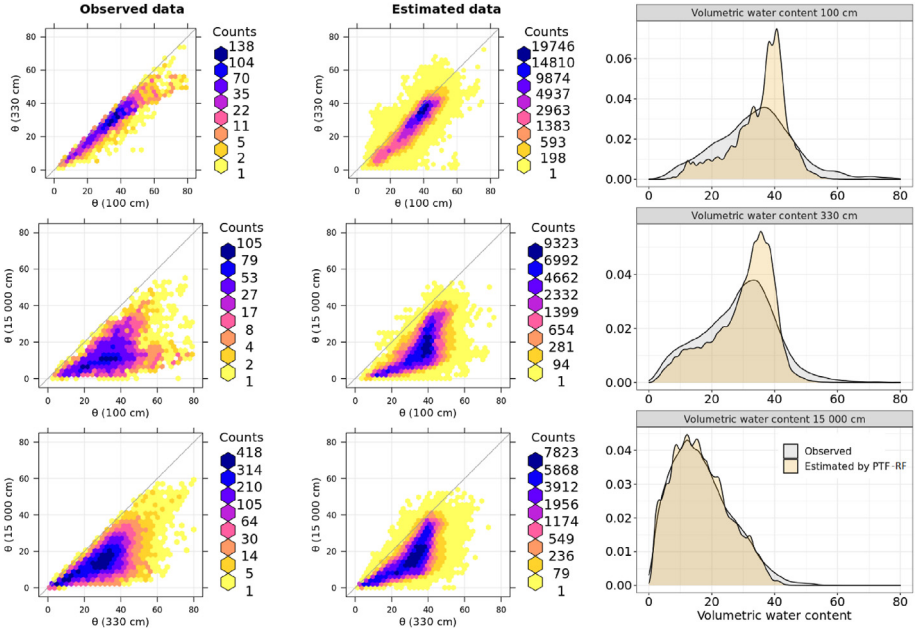

Fig. 3. Density scatter plots and density plots for volumetric water retention (100, 330 and 15 000 cm) represented in the screened dataset (observed) and the data from PTF-RFs (estimated). Volumetric water retention (θ) in 10^{-2} cm^3 cm^{-3}.

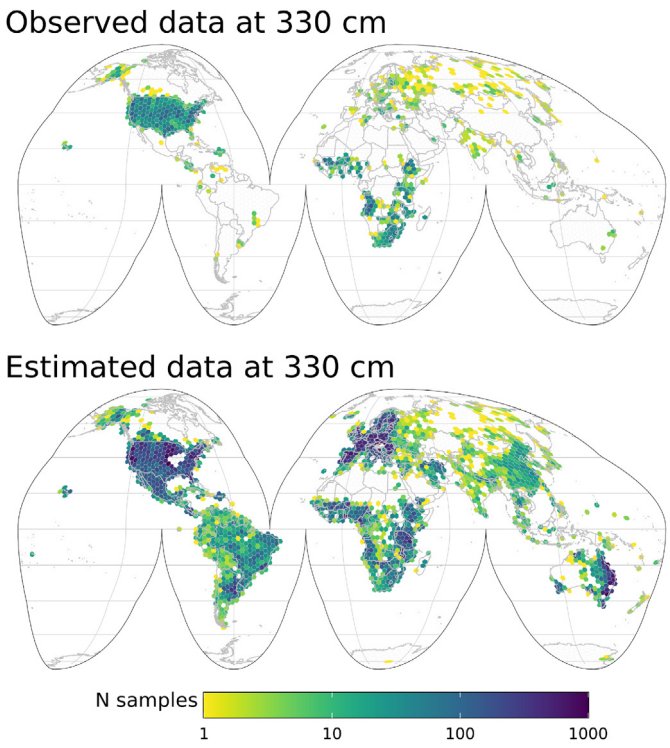

Fig. 4. Geographical distribution of volumetric water retention data at 330 cm after consistency evaluation (observed) and of data derived using PTF-RF application (estimated). Similar maps for the other pressures are provided in the Supplementary Materials (Section S3).

Table 6
Cross-validation statistics for evaluation of mapping performance using the complete dataset, a subset with only the observed data and a subset with only the estimated data. RMSE and ME in 10^{-2} cm^3 cm^{-3}.

	RMSE	ME	MEC
Volumetric soil water retention at 100 cm			
observed data	8.7	−0.8	0.440
estimated data	5.8	0.2	0.420
observed + estimated	6.4	0.1	0.430
Volumetric soil water retention at 330 cm			
observed data	7.6	0.2	0.345
estimated data	6.5	0.1	0.437
observed + estimated	7.1	0.2	0.386
Volumetric soil water retention at 15 000 cm			
observed data	6.6	0.2	0.494
estimated data	6.4	−0.1	0.464
observed + estimated	6.5	0.0	0.471

RMSE: root mean square error, ME: mean error, MEC: model efficiency coefficient.

and higher MEC. Similar patterns were observed by Dharumarajan et al. (2020), Malone et al. (2020), and Mashalaba et al. (2020). In general, this can be attributed to the lower variability of soil water retention data at higher suctions (Table 5).

Fig. 6 presents the results of the global mapping of soil water retention at 100, 330 and 15 000 cm for the depth interval of 0–5 cm. The geographical patterns were similar for 100 and 330 cm, especially in the northern hemisphere, possibly due to be relatively close suctions, while different patterns emerge at 15 000 cm. Results for the other depths are presented in the Supplementary Material (Section S5).

Fig. 7 presents the results for $q_{0.05}$, $q_{0.95}$ and PIR at 330 cm suction for 0–5 cm depth. For the presented suction and depth, high PIR values were predominant in South America and Australia, coinciding with areas that had a low observation density (Fig. 4). Alternatively, high values of PIR were also observed in the African continent, despite the relatively high observation density for that region. A possible explanation for this could be that the feature space is not covered as well in Africa as in other regions. Similar results are reported for the other suctions and depths (see Supplementary Materials, Section S6).

3.3. Comparison between point- and map-based mapping approaches

Map inter-comparison was performed based on scatter density plots and density plots of the data obtained at the hexagonal grid presented Fig. 8. The figure also shows the distribution of the volumetric water retention data. Irrespective of the soil water suction, Map3 and SoilGrids had the highest agreement. Map3 is based on the application of a VGM-PTF to basic soil property layers from the latest version of SoilGrids (Poggio et al., 2021), which may explain the similarities. Map1 and Map2 have multi-modal distributions with many values close to zero, for all three evaluated suctions, considering the grouped data from 0 to 200 cm Map1 in general has the highest degree of spatial variation, while Map2 has the lowest variation.

Fig. 9 shows the geographical distribution of soil water retention at 330 cm and 0–5 cm depth for the four maps. For the other suctions, we refer to the Supplementary Materials (Section S7). Patterns for the distribution of volumetric water retention at 100 cm for Map1, Map2 and Map3 were similar to those of the maps of saturated water retention (θ_S) presented in the original studies. The patterns and values of water retention on these maps were determined by the PTF that was used to build the maps and the base maps to which the PTF was applied (Table 1). Map1 has a more diverse geographical pattern than the other maps. This could in part be attributed to the fact that Dai, Xin, et al. (2019) used soil

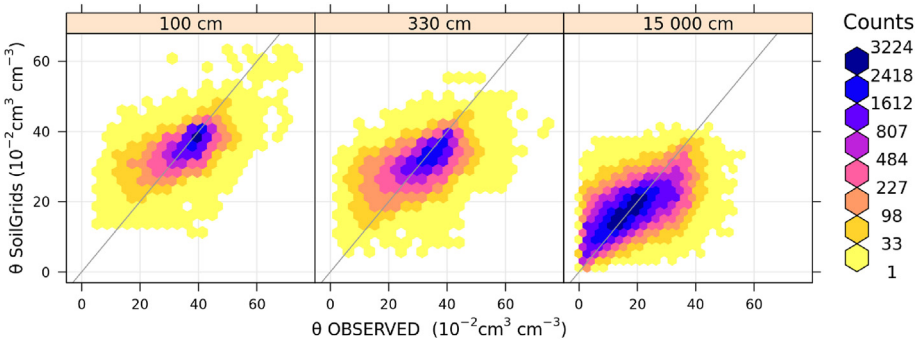

Fig. 5. Scatter density plots for SoilGrids cross-validation at all depth intervals considering the observed data.

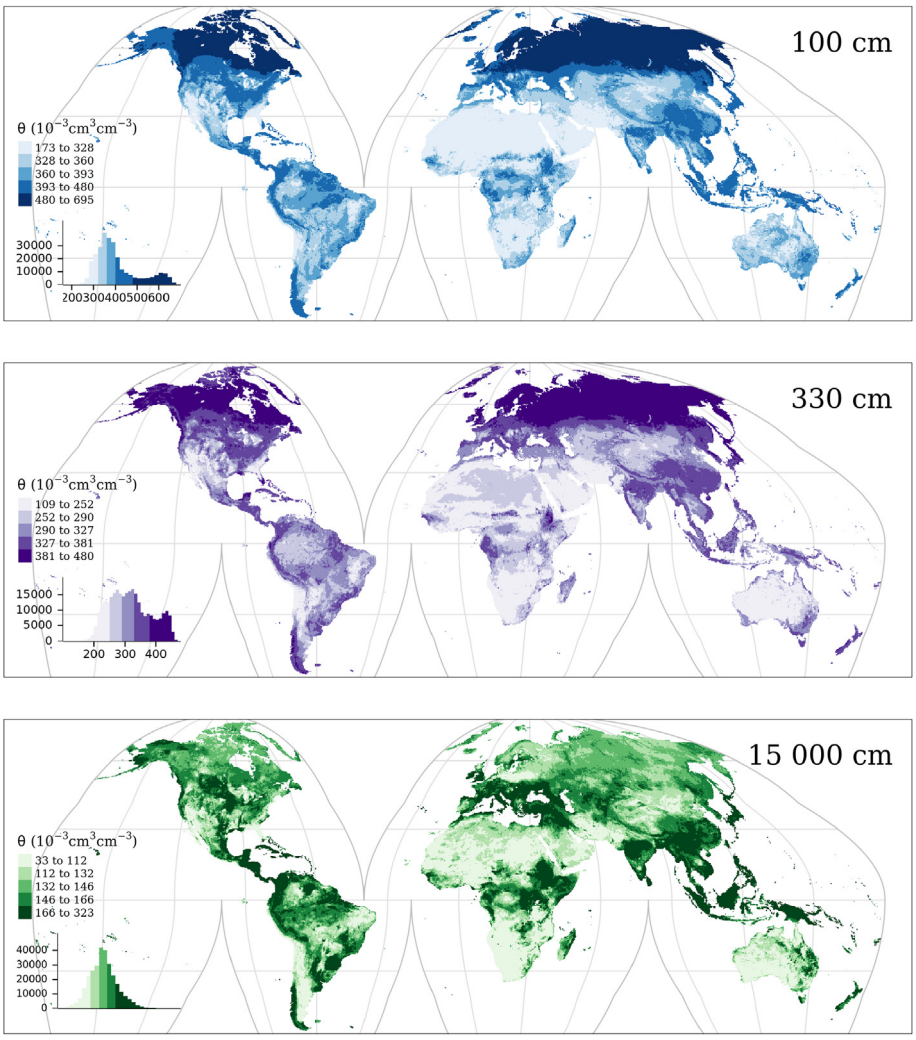

Fig. 6. Predicted mean volumetric soil water retention at 100 cm, 330 cm, and 15 000 cm for the 0–5 cm layer, including a global histogram of the mean soil water retention.

parameters based on the soil composition datasets of GSDE (Shangguan et al., 2014) applying various PTFs to describe soil water retention parameters.

3.4. Accuracy evaluation of point- and map-based mapping approaches

Point- and map-based derived maps were compared with the *observed* data to evaluate the prediction accuracy. Scatter density plots showing this evaluation are presented in the Supplementary Materials. Results for SoilGrids and Map3 appeared closer to the 1:1 line than for the other products (Fig. S34). Similar to the map comparison, Map2-predicted values had a narrower range than the other products. Conversely, Map1 presented very dispersed data relative to the 1:1 line. A possible explanation for this has been given in the previous section.

As indicated earlier, most observed data were located in the USA (Fig. 4). In the case of the observed data, the fraction corresponding to the USA was 49, 62 and 66% for 100, 330 and 15 000 cm, respectively, while for the estimated data the USA represented 56% of the total. To evaluate the effect of this unequal geographical distribution of observed data across the globe, Fig. 10 shows the RMSE of the maps for different sub datasets.

The four mapping approaches performed similarly when applied to the 'complete' and 'only USA' sub datasets, pointing to a high influence of these subsets on the final model evaluation. The mapping models for Map1 and Map2 are highly influenced by data from the USA, whereas this is less so for Map 3 and SoilGrids. However, the PTF used to build Map3 was developed considering European data only, and the apparent influence may be related to similarities between soils from temperate regions. Overall, the SoilGrids map and Map3 performed less well for the 'rest of the world' than for the complete and USA-only datasets. The general picture from Fig. 10 is that SoilGrids performed better than the three map-based derived maps, as reflected by lower RMSE values at the three suctions. Interestingly, for all maps, the RMSE for 100 cm was smaller than that for 330 cm and 15 000 cm.

It should be emphasized that the point-based comparison provides metrics to evaluate the quality of the maps at points. The point-based evaluation may not represent the accuracy of the maps when aggregated to larger supports, such as spatial means for fields or regions. It is, however, not possible to quantify the accuracy improvement due to 'upscaling' without modelling the spatial correlation of the prediction error (Webster & Oliver, 2007, pp. 153–194). Depending on the proposed applications, a more rigorous assessment of accuracies may be required.

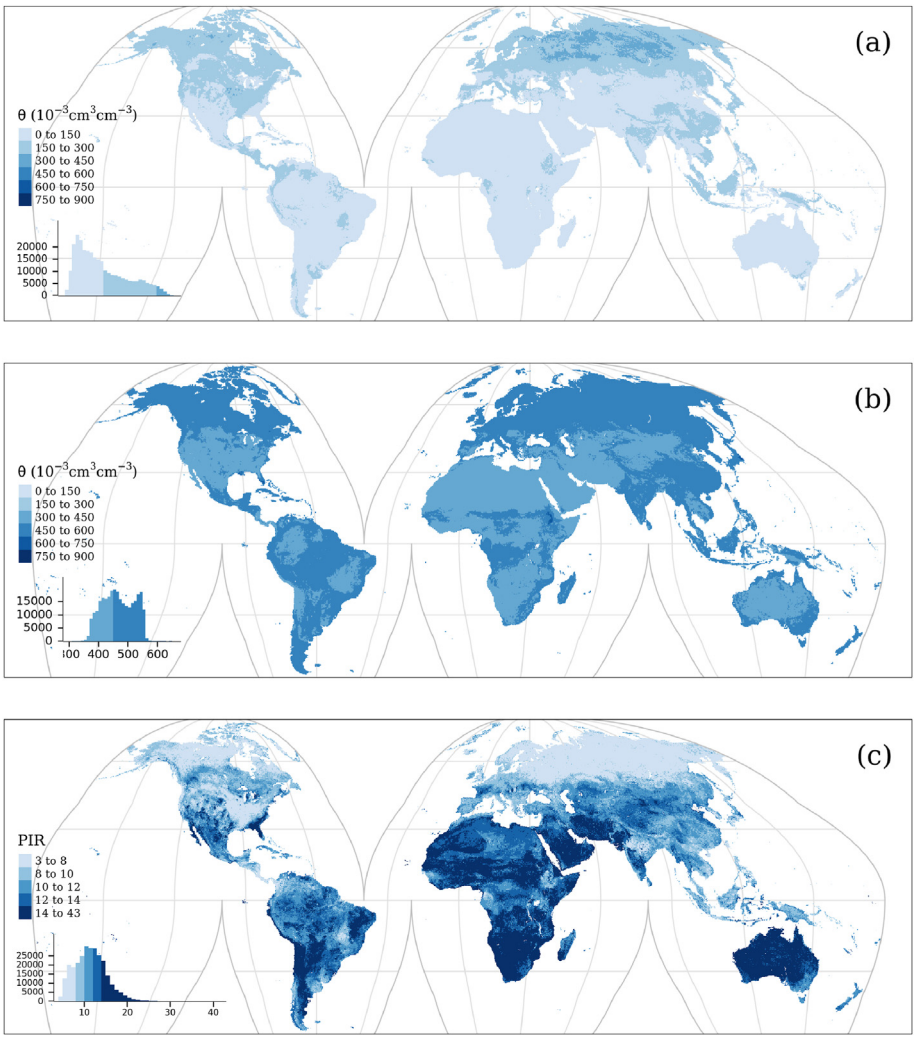

Fig. 7. Prediction distribution for soil water retention (θ) at 330 cm for the 0–5 cm depth interval: (A) 0.05 quantile, (B) 0.95 quantile and (C) the prediction interval ratio (PIR), including a global histogram.

3.5. General discussion

The quality of the point-based maps depends on the availability and consistency of the input data, the ability of the covariates to explain the spatial variation of soil water retention, and the performance of the mapping algorithm to capture the influence of covariates on the response variable. For the map-based approach, however, the quality is mainly determined by the PTF's capability to represent water retention in the evaluated area and the quality of the basic soil property maps, as indicated by Padarian et al. (2014).

Despite the undeniable functionality of PTFs, the assumptions for developing statistical PTFs imply a high degree of empiricism and uncertainty (Román Dobarco et al., 2019; Schaap & Leij, 1998; Vereecken et al., 2010). PTFs calibrated for one region may not be applicable in another region. As indicated by various authors (Dai, Xin, et al., 2019; Van Looy et al., 2017), PTF development must happen jointly with the development of appropriate extrapolation and upscaling techniques so that the PTFs can correctly represent the spatial heterogeneity of soils. As mentioned before, the use of a single PTF for global applications can lead to biases, underestimation of uncertainties, and overconfidence problems Dai, Xin, et al. (2019). In this study we used a single PTF-RF for enlarging the dataset used in the point-based mapping approach, as performed in other DSM projects (e.g. Hong et al. (2013); Liddicoat et al. (2015); Zare et al. (2021)), which also created uncertainties because PTF estimates are no substitute for real measurements. However, since our method also uses direct measurements the final results are likely to be less affected by these uncertainties, as confirmed by the cross-validation analysis. Nevertheless, one of the limitations of this work was to consider measured and PTF-RF-estimated data as equally important in the soil mapping approach. The various sources of uncertainty in the measurement themselves as well as in PTF-RF-estimated values, combined with those associated with the machine learning method, will lead to a higher uncertainty. Accuracy differences in calibration data will also affect their weight in the model calibration (e.g. Takoutsing et al. (2022); van der Westhuizen et al. (2022)). The effect of using such 'multi-source' point data should be evaluated in future research, as suggested in Wadoux et al. (2019).

The uncertainties and limitations of the point-based mapping method, as discussed by Poggio et al. (2021), include the limited prediction performance of the covariates, even though the set of potential covariates is continuously growing. In the case of soil water retention, the covariates are not able to represent all spatial variation. The variance explained by the model was 0.49 (Table 6), while for other soil/ecology attributes this may be up to 0.78

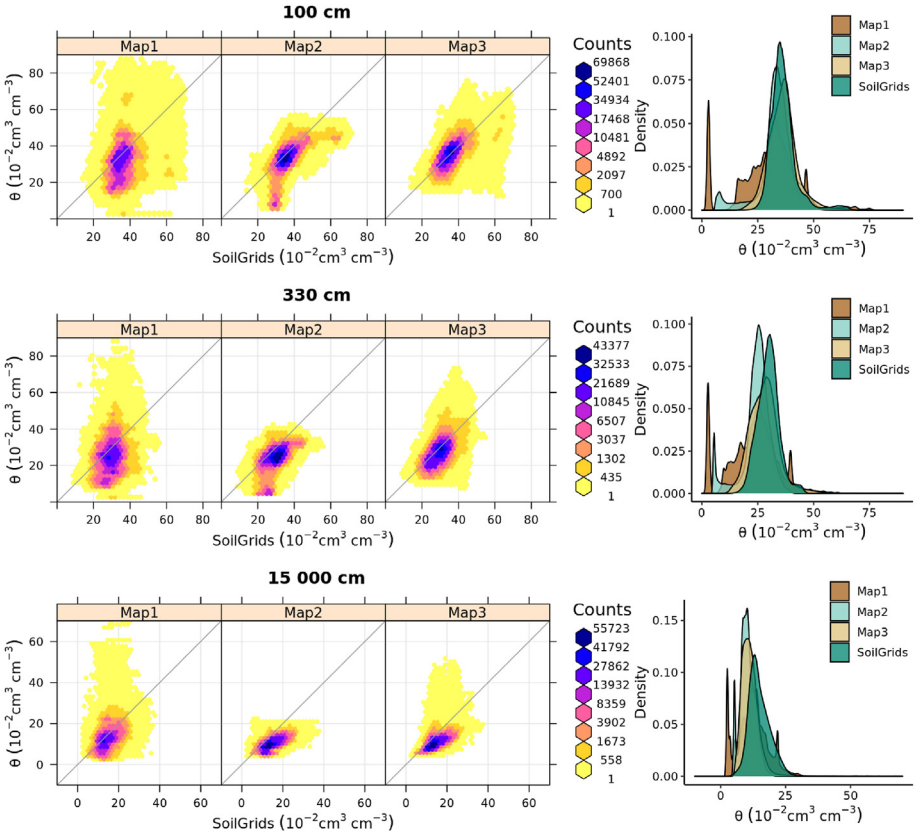

Fig. 8. Scatter density plots and density plots comparing the three selected maps with SoilGrids for points on a regular hexagonal grid at 100, 330 and 15 000 cm for all layers between 0 and 200 cm.

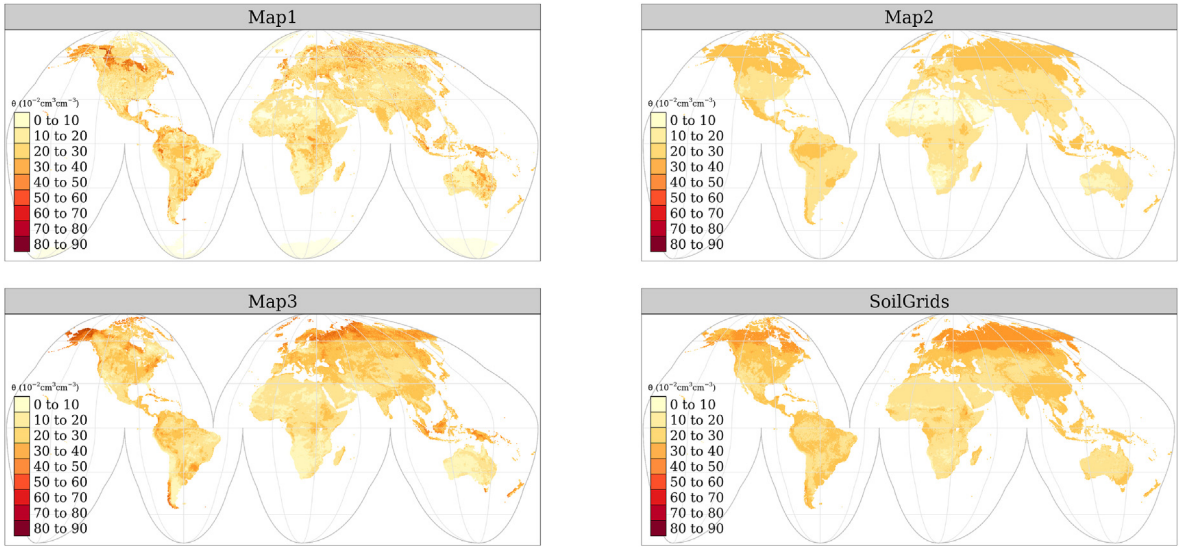

Fig. 9. Spatial variation in soil water retention observed between the evaluated maps at 330 cm for the 0–5 cm layer.

(Poggio et al., 2021).Selection of the DSM method used to model the data and generate maps also needs to be evaluated carefully. According to Zhang et al. (2017), considering the nature of the data and purposes of the mapping, random forest presents an up-to-date model with a fair compromise between performance and applicability.

It is also important to select an adequate method of cross-validation, since the use of clustered data for cross-validation tends to generate biased estimates of map accuracy (Wadoux et al., 2021). The influence of clustered data was observed in this study by the similar performances of all four maps considering the 'complete' and the 'USA only' datasets (Fig. 10). Since more than half of the data were from the USA, cross-validation results for the whole world ('complete dataset') were largely influenced by the

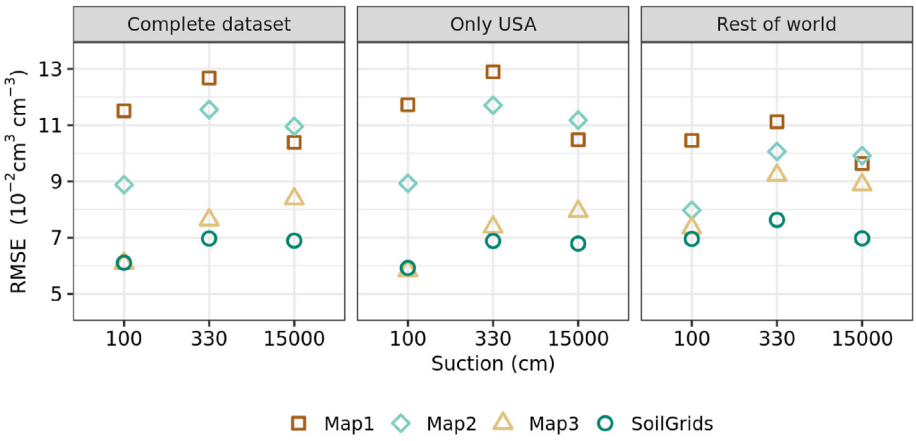

Fig. 10. Comparison of the RMSE of the evaluated maps at the cross-validation points considering all points at 100, 330 and 15 000 cm for the layer 0–200 cm.

performance in the USA. The cross-validation results for SoilGrids are similar to those reported in studies at a regional scale (Dharumarajan et al., 2020; Malone et al., 2020). The validation metrics were more optimistic, possibly due to the clustering.

The SoilGrids predictions were at point support, which implies large uncertainties because fine-scale spatial variation is large at point support and cannot be represented by medium-scale covariates. However, to quantify the uncertainty of spatial averages requires a geostatistical approach (Szatmári et al., 2021), which was beyond the scope of this study.

All the mentioned challenges of the point-based approach are also important for the map-based approach, since the generation of the basic soil property maps face the same challenges (Padarian et al., 2014). Despite the uncertainties associated with the present point-based mapping approach, the results of our study showed that point-based mapping reached a higher accuracy than map-based mapping, especially for 330 and 15 000 cm, for which there are more observations. Quantile maps (Fig. 7) and cross-validation statistics (Fig. 10) showed that prediction uncertainties are large, as also confirmed by the large differences between the three existing maps and SoilGrids. The spatial patterns obtained with the four mapping methods were also different (see Fig. 9 and Section S7 in the Supplementary Materials). The large RMSE (Fig. 10) and low MEC (Fig. S35) indicate that global maps should be treated with care. When high accuracy at finer resolution is desired, regional maps may be preferred (e.g. Dai et al. (2013); Malone et al. (2020)).

Further increase in accuracy and reduction of the uncertainty of global of soil water retention maps can be achieved by combining geographical databases of soil properties with remote sensing technology and proximal data sensing methods (Vereecken et al., 2016), using methods such as inverse modelling to derive soil water retention from temporal series of soil water moisture (Mohanty, 2013). Errors in the calibration data propagate and this can have a dramatic effect on subsequent modelling, so it would be even better if batch and laboratory measurement effects of the data stored in databases were routinely accompanied by measurement uncertainty metrics (van Leeuwen et al., 2021) so the influence of them in DSM is taken into account (van der Westhuizen et al., 2022), as well as other analytical and proximal soil sensing errors (Takoutsing et al., 2022).

Another important source of uncertainty that is not acknowledged in all evaluated maps is the temporal variation of soil water retention. For example, in the case of measurements taken in undisturbed soil samples, reorganisation of pore sizes due to tillage treatments can affect SWR at a given suction at point locations (Bescansa et al., 2006).

Despite all the past and present efforts to improve DSM methods and tools, one of the most important factors to determine the quality of the final map is the availability and quality of the input data. In this work, the creation of an appropriate dataset included several steps and filters to guarantee reasonable bounds and mutually consistent data. As a result, we discarded a fairly large proportion of the original data, in particular measurements on gravimetric basis where only 35, 39 and 27% of the available data at respectively 100, 330, and 15 000 cm were suitable for use, indicating that more effort should be made to quality-check data before they are added to a database. In this study, we used all available data from the WoSIS database and derived soil water retention maps with the methodology of first estimating data points with a PTF-RF and interpolating them later, which proved to be sufficient to reach a higher accuracy than with the map-based method of applying a PTF to basic soil maps. However, map accuracy would benefit greatly from an increase of calibration data, as well as improvements in attribute and positional accuracy (Batjes et al., 2020). Arrouays et al. (2020) emphasize the importance of acquiring harmonized, temporally varied data for developing maps which include the fluctuations of the variables in time. This would require the establishment of fairly detailed monitoring networks, and the subsequent sharing of the collected, geographically and temporally referenced data.

4. Conclusions

We evaluated the performance of point-based mapping of global soil water retention to predict data from the WoSIS database and compared the results with those of three map-based approaches. The comparison showed that the point-based derived maps (i.e. SoilGrids) performed better than the three map-based derived maps at 330 and 15 000 cm, yet with similar accuracy at 100 cm suction.

The uncertainty of the results of the point-based mapping was rather large which can be associated to two main factors: limited number of soil profile data available and limited prediction performance of covariates. In the case of soil water retention, the existing covariates were not able to capture all spatial variation. Hence the lower explained variance compared with similar approaches for mapping other soil properties or ecological variables. Uncertainty maps of the map-based methods were not available and therefore it was not possible to make a comparison for this.

Creating global maps of soil water retention through point-based mapping, with associated measures of prediction uncertainty, provides a promising approach for mapping hydraulic

properties for the world. Although the generated maps cannot be considered as directly obtained from point measurements because of the use of the PTF-RF to estimate a large amount of the input data, our results showed that this approach is potentially more suitable for global mapping of soil water retention. It better captures the spatial variation of soil water retention, and can result in increasingly accurate maps, provided an adequate number of evenly spatially-distributed soil samples become available. Further research to improve performance and reduce uncertainty would be beneficial, as well as integrating these products into global hydrological models.

Data availability

The SoilGrids layers are available for evaluation as pre-release at: WV 100 cm (https://doi.org/10.17027/isric-soilgrids.c6cb5073-78dd-4d8d-be81-9d546a1c004f); WV 330 cm (https://doi.org/10.17027/isric-soilgrids.14e7c761-6f87-4f4c-9035-adb282439a44); WV 15 000 cm (https://doi.org/10.17027/isric-soilgrids.f5a1188a-09f8-4ef6-b841-93f08e3903f4).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully thank the Editor and Reviewers for their valuable comments.

This work was undertaken as part of the Climate, Food and Farming and Global Research Alliance Development Scholarships (CLIFF-GRADS) program, an initiative implemented by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) and the Global Research Alliance on Agricultural Greenhouse Gases (GRA) with support from their donors. The actual work was carried out within the framework of the ISRIC guest researcher programme: MET is a PhD graduate from the Federal University of Paraná, Brazil.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.iswcr.2022.08.001.

References

Adhikari, K., & Hartemink, A. E. (2016). Linking soils to ecosystem services — a global review. *Geoderma, 262*, 101–111. https://doi.org/10.1016/j.geoderma.2015.08.009

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., Mendonca-Santos, M.d. L., Minasny, B., Montanarella, L., Odeh, I. O., Sanchez, P. A., Thompson, J. A., & Zhang, G. L. (2014). *Chapter three - globalsoilmap: Toward a fine-resolution global grid of soil properties, 125* pp. 93–134). Academic Press. https://doi.org/10.1016/B978-0-12-800137-0.00003-0. of Advances in Agronomy.

Arrouays, D., Poggio, L., Salazar Guerrero, O. A., & Mulder, V. L. (2020). Digital soil mapping and globalsoilmap. main advances and ways forward. *Geoderma Regional, 21*, Article e00265. https://doi.org/10.1016/j.geodrs.2020.e00265

Barnes, R., Sahr, K., Evenden, G., Johnson, A., Warmerdam, F., Rouault, E., & Song, L. (2020). dggridR: Discrete global grids for R. URL: https://github.com/r-barnes/dggridR/. r package version 2.0.4.

Batjes, N. H., Ribeiro, E., & van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (wosis snapshot 2019). *Earth System Science Data, 12*, 299–320. https://doi.org/10.5194/essd-12-299-2020

Bescansa, P., Imaz, M., Virto, I., Enrique, A., & Hoogmoed, W. (2006). Soil water retention as affected by tillage and residue management in semiarid Spain. *Soil and Tillage Research, 87*, 19–27. https://doi.org/10.1016/j.still.2005.02.028

Bivand, R., Keitt, T., & Rowlingson, B. (2021). rgdal: Bindings for the 'Geospatial' data abstraction library. URL: https://CRAN.R-project.org/package=rgdal. r package version 1.5-23.

Blake, G. R. (2008). *Particle density*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-3995-9_406

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Campbell, G. S. (1974). A simple method for determining unsaturated conductivity from moisture retention data. *Soil Science, 117*, 311–314. https://doi.org/10.1097/00010694-197406000-00001

Dai, Y., Shangguan, W., Duan, Q., Liu, B., Fu, S., & Niu, G. (2013). Development of a China dataset of soil hydraulic parameters using pedotransfer functions for land surface modeling. *Journal of Hydrometeorology, 14*, 869–887. https://doi.org/10.1175/JHM-D-12-0149.1

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., & Yan, F. (2019). A review of the global soil property maps for earth system models. *SOIL, 5*, 137–158. https://doi.org/10.5194/soil-5-137-2019

Dai, Y., Xin, Q., Wei, N., Zhang, Y., Shangguan, W., Yuan, H., Zhang, S., Liu, S., & Lu, X. (2019). A global high-resolution data set of soil hydraulic and thermal properties for land surface modeling. *Journal of Advances in Modeling Earth Systems, 11*, 2996–3023. https://doi.org/10.1029/2019MS001784

De Lannoy, G. J. M., Koster, R. D., Reichle, R. H., Mahanama, S. P. P., & Liu, Q. (2014). An updated treatment of soil texture and associated hydraulic properties in a global land modeling system. *Journal of Advances in Modeling Earth Systems, 6*, 957–979. https://doi.org/10.1002/2014MS000330

Dharumarajan, S., Kalaiselvi, B., Suputhra, A., Lalitha, M., Hegde, R., Singh, S., & Lagacherie, P. (2020). Digital soil mapping of key globalsoilmap properties in northern Karnataka plateau. *Geoderma Regional, 20*, Article e00250. https://doi.org/10.1016/j.geodrs.2019.e00250

Dobriyal, P., Qureshi, A., Badola, R., & Hussain, S. A. (2012). A review of the methods available for estimating soil moisture and its implications for water resource management. *Journal of Hydrology, 458–459*, 110–117. https://doi.org/10.1016/j.jhydrol.2012.06.021

Eitzinger, J., Trnka, M., Hösch, J., Zalud, Z., & Dubrovský, M. (2004). Comparison of ceres, wofost and swap models in simulating soil water content during growing season under different soil conditions. *Ecological Modelling, 171*, 223–246. https://doi.org/10.1016/j.ecolmodel.2003.08.012

Han, E., Ines, A. V., & Koo, J. (2019). Development of a 10-km resolution global soil profile dataset for crop modeling applications. *Environmental Modelling & Software, 119*, 70–83. https://doi.org/10.1016/j.envsoft.2019.05.012

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G. B., Walsh, M. G., & González, R. (2014). Soilgrids1km - global soil information based on automated mapping. *PLoS One, 9*, 1–17. https://doi.org/10.1371/journal.pone.0105992

Heuscher, S. A., Brandt, C. C., & Jardine, P. M. (2005). Using soil physical and chemical properties to estimate bulk density. *Soil Science Society of America Journal, 69*, 51–56. https://doi.org/10.2136/sssaj2005.0051a

Hijmans, R. J. (2021a). raster: Geographic data analysis and modeling. URL: https://CRAN.R-project.org/package=raster. r package version 3.4-10.

Hijmans, R. J. (2021b). terra: Spatial data analysis. URL: https://CRAN.R-project.org/package=terra. r package version 1.2-10.

Hong, S. Y., Minasny, B., Han, K. H., Kim, Y., & Lee, K. (2013). Predicting and mapping soil available water capacity in korea. *PeerJ, 1*, e71. https://doi.org/10.7717/peerj.71

Hoogenboom, G., Wilkens, P. W., & Tsuji, G. Y. (Eds.). (1999). *DSSAT v3, ume 4*. Honolulu, Hawaii: University of Hawaii. URL: https://dssat.net/wp-content/uploads/2011/10/DSSAT-vol4.pdf.

Janssen, P., & Heuberger, P. (1995). Calibration of process-oriented models. *Ecological Modelling, 83*, 55–66. https://doi.org/10.1016/0304-3800(95)00084-9. modelling Water, Carbon and Nutrient Cycles in Forests.

de Jong van Lier, Q., Pinheiro, E. A. R., & Inforsato, L. (2018). A one-dimensional physically based approach to predict soil profile aeration requirements. *Soil Science Society of America Journal, 82*, 593–600. https://doi.org/10.2136/sssaj2017.10.0369

Kuhn, M. (2021). caret: Classification and regression training. URL: https://CRAN.R-project.org/package=caret. r package version 6.0-88.

Laio, F., Porporato, A., Ridolfi, L., & Rodriguez-Iturbe, I. (2001). Plants in water-controlled ecosystems: Active role in hydrologic processes and response to water stress: Ii. Probabilistic soil moisture dynamics. *Advances in Water Resources, 24*, 707–723. https://doi.org/10.1016/S0309-1708(01)00005-7

Leenaars, J. G., Claessens, L., Heuvelink, G. B., Hengl, T., Ruiperez González, M., van Bussel, L. G., Guilpart, N., Yang, H., & Cassman, K. G. (2018). Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-saharan africa. *Geoderma, 324*, 18–36. https://doi.org/10.1016/j.geoderma.2018.02.046

van Leeuwen, C. C. E., Mulder, V. L., Batjes, N. H., & Heuvelink, G. B. M. (2021). Statistical modelling of measurement error in wet chemistry soil data. *European Journal of Soil Science*. https://doi.org/10.1111/ejss.13137

Liddicoat, C., Maschmedt, D., Clifford, D., Searle, R., Herrmann, T., Macdonald, L. M., & Baldock, J. (2015). Predictive mapping of soil organic carbon stocks in south Australia's agricultural zone. *Soil Research, 53*, 956–973. https://doi.org/10.1071/SR15100

Malone, B., Luo, Z., He, D., Viscarra Rossel, R., & Wang, E. (2020). Bioclimatic variables as important spatial predictors of soil hydraulic properties across Australia's agricultural region. *Geoderma Regional, 23*, Article e00344. https://doi.org/10.1016/j.geodrs.2020.e00344

Mashalaba, L., Galleguillos, M., Seguel, O., & Poblete-Olivares, J. (2020). Predicting spatial variability of selected soil properties using digital soil mapping in a rainfed vineyard of central Chile. *Geoderma Regional, 22*, Article e00289. https://doi.org/10.1016/j.geodrs.2020.e00289

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution, 12*, 1620–1633. https://doi.org/10.1111/2041-210X.13650

Minasny, B., Hopmans, J. W., Harter, T., Eching, S. O., Tuli, A., & Denton, M. A. (2004). Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. *Soil Science Society of America Journal, 68*, 417–429. https://doi.org/10.2136/sssaj2004.4170

Mohanty, B. P. (2013). Soil hydraulic property estimation using remote sensing: A review. *Vadose Zone Journal, 12*. https://doi.org/10.2136/vzj2013.06.0100. vzj2013.06.0100.

Montzka, C., Herbst, M., Weihermüuller, L., Verhoef, A., & Vereecken, H. (2017). A global data set of soil hydraulic properties and sub-grid variability of soil water retention and hydraulic conductivity curves. *Earth System Science Data, 9*, 529–543. https://doi.org/10.5194/essd-9-529-2017

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research, 12*, 513–522. https://doi.org/10.1029/WR012i003p00513

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology, 10*, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nemes, A., Quebedeaux, B., & Timlin, D. J. (2010). Ensemble approach to provide uncertainty estimates of soil bulk density. *Soil Science Society of America Journal, 74*, 1938–1945. https://doi.org/10.2136/sssaj2009.0370

Nemes, A., Schaap, M., Leij, F., & Wösten, J. (2001). Description of the unsaturated soil hydraulic database unsoda version 2.0. *Journal of Hydrology, 251*, 151–162. https://doi.org/10.1016/S0022-1694(01)00465-6

Padarian, J., Minasny, B., McBratney, A., & Dalgliesh, N. (2014). Predicting and mapping the soil available water capacity of australian wheatbelt. *Geoderma Regional, 2*(3), 110–118. https://doi.org/10.1016/j.geodrs.2014.09.005

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal, 10*, 439–446. https://doi.org/10.32614/RJ-2018-009. URL:.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). Soilgrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL, 7*, 217–240. https://doi.org/10.5194/soil-7-217-2021

Poggio, L., Gimona, A., Brown, I., & Castellazzi, M. (2010). Soil available water capacity interpolation and spatial uncertainty modelling at multiple geographical extents. *Geoderma, 160*, 175–188. https://doi.org/10.1016/j.geoderma.2010.09.015

Porporato, A., Feng, X., Manzoni, S., Mau, Y., Parolari, A. J., & Vico, G. (2015). Ecohydrological modeling in agroecosystems: Examples and challenges. *Water Resources Research, 51*, 5081–5099. https://doi.org/10.1002/2015WR017289

Pumo, D., Viola, F., & Noto, L. V. (2008). Ecohydrology in mediterranean areas: A numerical model to describe growing seasons out of phase with precipitations. *Hydrology and Earth System Sciences, 12*, 303–316. https://doi.org/10.5194/hess-12-303-2008

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Raes, D., Steduto, P., Hsiao, T. C., & Fereres, E. (2016). *Chapter 1 FAO crop-water productivity model to simulate yield response to water*. Rome, Italy: FAO. URL: http://www.fao.org/documents/card/en/c/BR246E.

Reynolds, C. A., Jackson, T. J., & Rawls, W. J. (2000). Estimating soil water-holding capacities by linking the food and agriculture organization soil map of the world with global pedon databases and continuous pedotransfer functions. *Water Resources Research, 36*, 3653–3662. https://doi.org/10.1029/2000WR900130

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

Saxton, K. E., & Rawls, W. J. (2006). Soil water characteristic estimates by texture and organic matter for hydrologic solutions. *Soil Science Society of America Journal, 70*, 1569–1578. https://doi.org/10.2136/sssaj2005.0117

Saxton, K. E., Rawls, W. J., Romberger, J. S., & Papendick, R. I. (1986). Estimating generalized soil-water characteristics from texture. *Soil Science Society of America Journal, 50*, 1031–1036. https://doi.org/10.2136/sssaj1986.03615995005000040039x

Schaap, M. G., & Leij, F. J. (1998). Database-related accuracy and uncertainty of pedotransfer functions. *Soil Science, 163*, 765–779.

Schaap, M. G., Leij, F. J., & van Genuchten, M. T. (2001). Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology, 251*, 163–176. https://doi.org/10.1016/S0022-1694(01)00466-8

Sequeira, C. H., Wills, S. A., Seybold, C. A., & West, L. T. (2014). Predicting soil bulk density for incomplete databases. *Geoderma, 213*, 64–73. https://doi.org/10.1016/j.geoderma.2013.07.013

Seybold, C. A., Harms, D. S., & Williams, C. O. (2014). Soil survey: Prediction of bulk density using k-nearest neighbor approach. *Soil Horizons*, 1–11. https://doi.org/10.2136/sh13-05-0014

Shangguan, W., Dai, Y., Duan, Q., Liu, B., & Yuan, H. (2014). A global soil data set for earth system modeling. *Journal of Advances in Modeling Earth Systems, 6*, 249–263. https://doi.org/10.1002/2013MS000293

Simons, G., Koster, R., & Droogers, P. (2020). *HiHydroSoil v2.0 - a high resolution soil map of global hydraulic properties*. Technical Report. FutureWater report 213. Wageningen, The Netherlands.

Soil Survey Staff, U. (2014). *Kellogg soil Survey laboratory methods manual. Soil Survey investigations Report No. 42, version 5.0*. Department of Agriculture, Natural Resources Conservation Service. Technical Report.

de Sousa, L. M., Poggio, L., Dawes, G., Kempen, B., & van den Bosch, R. (2020). Computational infrastructure of soilgrids 2.0. In I. N. Athanasiadis, S. P. Frysinger, G. Schimak, & W. J. Knibbe (Eds.), *Environmental software systems. Data science in action* (pp. 24–31). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-39815-6_3.

de Souza, E., Fernandes Filho, E. I., Schaefer, C. E. G. R., Batjes, N. H., Santos, G. R.d., & Pontes, L. M. (2016). Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio doce basin. *Scientia Agricola, 73*, 525–534. https://doi.org/10.1590/0103-9016-2015-0485

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., & Pásztor, L. (2019). Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics. *Hydrology and Earth System Sciences, 23*, 2615–2635. https://doi.org/10.5194/hess-23-2615-2019

Szatmári, G., Pásztor, L., & Heuvelink, G. B. M. (2021). Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. *Geoderma, 403*, Article 115356. https://doi.org/10.1016/j.geoderma.2021.115356. URL: https://www.sciencedirect.com/science/article/pii/S0016706121004365.

Takoutsing, B., Heuvelink, G. B. M., Stoorvogel, J. J., Shepherd, K. D., & Aynekulu, E. (2022). Accounting for analytical and proximal soil sensing errors in digital soil mapping. *European Journal of Soil Science, 73*, Article e13226. https://doi.org/10.1111/ejss.13226

Tennekes, M. (2018). tmap: Thematic maps in R. *Journal of Statistical Software, 84*, 1–39. https://doi.org/10.18637/jss.v084.i06

Tennekes, M. (2021). tmaptools: Thematic map tools. URL: https://CRAN.R-project.org/package=tmaptools. r package version 3.1-1.

Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G., & Tóth, G. (2015). New generation of hydraulic pedotransfer functions for europe. *European Journal of Soil Science, 66*, 226–238. https://doi.org/10.1111/ejss.12192

Twarakavi, N. K. C., Sakai, M., & Šimůnek, J. (2009). An objective analysis of the dynamic nature of field capacity. *Water Resources Research, 45*, Article W10410. https://doi.org/10.1029/2009WR007944

Van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal, 44*, 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

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., & Vereecken, H. (2017). Pedotransfer functions in earth system science: Challenges and perspectives. *Reviews of Geophysics, 55*, 1199–1256. https://doi.org/10.1002/2017RG000581

Vasques, G. M., Coelho, M. R., Dart, R. O., Oliveira, R. P., & Teixeira, W. G. (2016). Mapping soil carbon, particle-size fractions, and water retention in tropical dry forest in Brazil. *Pesquisa Agropecuária Brasileira, 51*, 1371–1385. https://doi.org/10.1590/s0100-204x2016000900036

Vereecken, H., Huisman, J. A., Bogena, H., Vanderborght, J., Vrugt, J. A., & Hopmans, J. W. (2008). On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research, 44*. https://doi.org/10.1029/2008WR006829

Vereecken, H., Schnepf, A., Hopmans, J., Javaux, M., Or, D., Roose, T., Vanderborght, J., Young, M., Amelung, W., Aitkenhead, M., Allison, S., Assouline, S., Baveye, P., Berli, M., Brüggemann, N., Finke, P., Flury, M., Gaiser, T., Govers, G., … Young, I. (2016). Modeling soil processes: Review, key challenges, and new perspectives. *Vadose Zone Journal, 15*. https://doi.org/10.2136/vzj2015.09.0131

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M. G., & van Genuchten, M. (2010). Using pedotransfer functions to estimate the van genuchten-mualem soil hydraulic properties: A review. *Vadose Zone Journal, 9*, 795–820. https://doi.org/10.2136/vzj2010.0045

Wadoux, A. M. C., Heuvelink, G. B., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling, 457*, Article 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692

Wadoux, A. M. J. C., Padarian, J., & Minasny, B. (2019). Multi-source data integration for soil mapping using deep learning. *SOIL, 5*, 107–119. https://doi.org/10.5194/soil-5-107-2019

Webster, R., & Oliver, M. (2007). *Local estimation or prediction: Kriging*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470517277.ch8 (chapter 8).

van der Westhuizen, S., Heuvelink, G. B., Hofmeyr, D. P., & Poggio, L. (2022). Measurement error-filtered machine learning in digital soil mapping. *Spatial Statistics, 47*, Article 100572. https://doi.org/10.1016/j.spasta.2021.100572

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*, 1686. https://doi.org/10.21105/joss.01686

Wösten, J., Pachepsky, Y., & Rawls, W. (2001). Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology, 251*, 123–150. https://doi.org/10.1016/S0022-1694(01)

00464-4

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*, 1–17. https://doi.org/10.18637/jss.v077.i01

Zare, S., Abtahi, A., Fallah Shamsi, S. R., & Lagacherie, P. (2021). Combining laboratory measurements and proximal soil sensing data in digital soil mapping approaches. *Catena, 207*, Article 105702. https://doi.org/10.1016/j.catena.2021.105702. URL: https://www.sciencedirect.com/science/article/pii/S0341816221005609.

Zhang, G.l., Liu, F., & Song, X.d. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture, 16*, 2871–2885. https://doi.org/10.1016/S2095-3119(17)61762-3

Zhang, Y., Schaap, M. G., & Wei, Z. (2020). Development of hierarchical ensemble model and estimates of soil water retention with global coverage. *Geophysical Research Letters, 47*, Article e2020GL088819. https://doi.org/10.1029/2020GL088819

Zhang, Y., Schaap, M. G., & Zha, Y. (2018). A high-resolution global map of soil hydraulic properties produced by a hierarchical parameterization of a physically based water retention model. *Water Resources Research, 54*, 9774–9790. https://doi.org/10.1029/2018WR023539